RICE UNIVERSITY

Nonlinear Signal Models: Geometry, Algorithms, and Analysis

by

Chinmay Hegde

A Thesis Submitted
in Partial Fulfillment of the
Requirements for the Degree

Doctor of Philosophy

Approved, Thesis Committee:

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

Dr. Ashok Veeraraghavan
Assistant Professor
Electrical and Computer Engineering

Dr. Wotao Yin
Associate Professor
Computational and Applied Mathematics

Houston, Texas

September 2012

Abstract

Traditional signal processing systems, based on linear modeling principles, face a stifling pressure to meet present-day demands caused by the deluge of data generated, transmitted and processed across the globe. Fortunately, recent advances have resulted in the emergence of more sophisticated, nonlinear signal models. Such nonlinear models have inspired fundamental changes in which information processing systems are designed and analyzed. For example, the sparse signal model serves as the basis for Compressive Sensing (CS), an exciting new framework for signal acquisition.

In this thesis, we advocate a geometry-based approach for nonlinear modeling of signal ensembles. We make the guiding assumption that the signal class of interest forms a nonlinear low-dimensional manifold belonging to the high-dimensional signal space. A host of traditional nonlinear data models can be essentially interpreted as specific instances of such manifolds. Therefore, our proposed geometric approach provides a common framework that can unify, analyze, and significantly extend the scope of nonlinear models for information acquisition and processing.

We demonstrate that the geometric approach enables new algorithms and analysis for a number of signal processing applications. Our specific contributions include: (i) new convex formulations and algorithms for the design of linear systems for data acquisition, compression, and classification; (ii) a general algorithm for reconstruction, deconvolution, and denoising of signals, images, and matrix-valued data; (iii) efficient methods for inference from a small number of linear signal samples, without ever resorting to reconstruction; and, (iv) new signal and image representations for robust modeling and processing of large-scale data ensembles.

Acknowledgements

This thesis is the culmination of six memorable years spent in graduate school at Rice and I would be remiss not to thank all the wonderful people that have contributed, directly or otherwise, to its creation.

First and foremost, I wish to express my gratitude to my inimitable advisor, Rich Baraniuk. Rich's unfettered joy and enthusiasm for research, and for life in general, has been a constant source of inspiration. I am fortunate to have been his student and thank him for the countless hours that he has patiently spent trying to mold me into a better researcher, writer, and person. I will also fondly remember our efforts in home improvement, slide-making all nighters, bio photo experiments, the ICASSP challenge, and much more.

I wish to express my gratitude to my thesis committee, Ashok Veeraraghavan and Wotao Yin, for their great perspective and feedback. My interaction with Ashok started during a terrific summer internship at Mitsubishi Electric Research Labs (MERL) and has persisted ever since. I thank him for his boundless energy and his extremely incisive questions. I also thank Wotao for his great patience and kindness while providing feedback to my ideas, and for introducing me to convex optimization when I started off at Rice.

I have been fortunate to work with a large and talented group of people whose contributions are reflected throughout this thesis: Rich Baraniuk, Volkan Cevher, Mark Davenport, Marco Duarte, Devin Grady, Piotr Indyk, Lydia Kavraki, Kevin Kelly, Jason Laska, Mark Moll, Yun Li, Sriram Nagraj, Fatih Porikli, Aswin Sankaranarayanan, Stephen Schnelle, Oncel Tuzel, Mike Wakin, and Wotao Yin. I thank each of them for helping me grow as a researcher.

My academic experience at Rice has been enriched by a motley crew of students, post-docs, professors, visitors, and researchers sometimes known as the 'DSP Group' (exact definition still unknown). I wish to express my thanks to Ali A., Ali M., Amirali, Arian, Aswin, Christoph, Drew, Denise, Eva, Jarvis, Jason, Jianing, JP, Laurent, Lee, Liz, Manjari, Marco, Mark, Matthew, Mike, Mona, Mr. Lan, Mr. Tan, Petros, Piotr, Raajen, Rich, Shriram, Sriram, Stephen, Volkan, and many others. I could not have asked for a more interesting group to work with. Special thanks to my

office mates Chris S., Drew, Eva, Marco, and Mark; you guys made me eagerly look forward to coming in to work every day.

Outside of work, I have been fortunate to share several fun-filled experiences with a large circle of friends at Rice, including Abhilash, Abhinav, Achal, Aparna, Animesh, Anubha, Arjun, Avani, Chaitra, Charu, Deepa, Dharma, Gautam, Harshit, Jatin, Kaushik, Kishori, Nikhil, Raghavan, Ramdas, Ramya, Richa, Sayantan, Shyam, Sravani, Sumedh, Vijetha, Vishnu, and many more. I will always remember the samosas, the music, and the campfires. Special thanks to my roommates Abhilash and Arjun for their valiant efforts to teach me how to cook, and to Achal for providing several memorable moments of enlightenment, both within Duncan Hall and without. Thanks also to my friends scattered across the globe: Aaftaab, Auditya, Avinash, Bala, Hari, Kamesh, RK, Supradeep, Shaurjo, Shreekrishna, and Sundeep, for the great memories. I hope that they continue to find happiness in wherever their paths lead them.

Lastly, I wish to thank a special group of people that are the real reasons for my continued good fortune: my parents, Shri Shantaram and Smt. Jaya, my grandmother Smt. Lalita, and my lovely girlfriend Deepti. Any success of mine would be impossible without your constant love and encouragement.

Contents

List of Figures

List of Algorithms

CHAPTER 1

Introduction

1.1 Signal Processing: Goals and Impact

We live in the *Age of Information*. The average American consumes upwards of 35 gigabytes worth of information every day [1]. This corpus of information presents itself in a plethora of forms, including audio, images, video, and text, from a plethora of sources, including television, radio, newspapers, magazines, and the World Wide Web. The growth in the availability of reliable, meaningful information is arguably the biggest driver of the transformations that modern society has witnessed over the last two decades, with dramatic, tangible impacts in diverse fields such as physics, medicine, finance, and entertainment.

The exponential increase in available information raises the specter of a *data deluge* [2–4]. As an example, a typical oil and gas company generates upwards of 10 gigabytes of data per day from each of several thousands of digital telemetry sensors deployed at its oil-fields [5]. As another example, 200 of London's traffic cameras generate and transmit over 8 terabytes of video data per day to a central monitoring unit [6]. The management of information volumes at such magnitudes poses significant scientific and engineering challenges.

Signal processing can be dubbed, in very general terms, as the *science of information*. Core principles in signal processing research encompass the design and analysis of: efficient sensors that record physical phenomena; numerical methods that transform the data into a computationally malleable form; systems that permit tractable data manipulation and transmission; and algorithms that extract meaningful, relevant information from raw data. Nevertheless, current information systems face a stifling pressure to meet the present-day demands generated by the data deluge and traditional signal processing principles do not always have adequate answers for present-day challenges. As a consequence, signal processing research is now at a crossroads.

1.2 Models in Signal Processing

We focus our attention on the notion of a *signal model*. Informally, a model is a mathematical descriptions that distinguishes, among a given class of signals, the interesting signals from the rest. There is often reason to believe that interesting signals are governed by only a few degrees of freedom, or a small number of physically meaningful parameters, relative to their size. In other words, interesting signals exhibit some notion of *conciseness* or *structure*.

Signal models are formal mathematical constructions that capture this belief of conciseness among the signals of interest. The core premise is that such a concise mathematical description can enable an improved understanding of the problem at hand, as well as result in efficient numerical algorithms. The notion of a signal model forms the bedrock of every signal processing principle, sometimes subtly so. For example, the celebrated Nyquist Sampling Theorem [7–10] specifies that a continuous-time signal with a finite bandwidth can be exactly reconstructed from uniform discrete samples at a rate faster than twice its bandwidth. Here, the key assumption is that the signals of interest are limited in bandwidth, and therefore the underlying (conceptual) signal model captures the notion of bandlimitedness.

The vast majority of traditional techniques in signal and image processing, communications, and control are based on *linear models*. Such models reflect the conceptual belief that a linear combination of two signals also gives rise to a physically meaningful signal[1]. A number of core signal processing principles can be interpreted and analyzed in terms of linear models. For example, in the case of the Nyquist Sampling Theorem, the underlying bandlimitedness assumption, in fact, represents a linear model. This is easily observed due to the fact that the linear sum of signals of bandwidth W is also a signal of bandwidth W.

From a mathematical perspective, linear signal models are conveniently characterized via the geometric notion of a *subspace* of an ambient, high-dimensional signal space. Traditional signal processing techniques such as signal sampling, reconstruction, filtering, and processing can be analyzed via operators acting upon linear subspaces [11]. Subspace-based approaches have also been successfully applied to other signal processing applications such as estimation, tracking, detection, and denoising [12]. The linear modeling approach not only enables efficient analysis, but also affects system design and implementation. Large, complex digital and analog signal processing systems are constructed using simple building blocks such as filters, which are usually modeled as linear systems [11]. Additionally, a number of important control systems are designed to operate in the linear regime [13].

1.3 Nonlinear Models

In several application domains, the linear assumption is only a poor approximation. A simple conceptual instance is the processing of real-world images. The linear superposition of two real-world images rarely produces a realistic-looking image. Indeed, conventional methods such as linear filtering of a noisy image often results in significant blurring and/or other artifacts. Therefore, contrary to traditional linear models, a parallel approach is to develop *nonlinear* models and methods to

[1] In physics, this is more popularly known as the *principle of superposition.*

design and analyze signal processing systems.

For the most part, the development of core signal processing principles has excluded nonlinear models. In typical problems, it is not uncommon to make a linearizing assumption, even though such an approximation is known to be poor. The emphasis on linear models and methods is unsurprising. Nonlinear models and methods are often marked by concerns marking their *numerical stability*; nonlinear algorithms typically incur great *computational costs*; and nonlinear approaches are *difficult to characterize* from an analytical standpoint.

Somewhat strikingly therefore, over the last decade, there have been impressive advances in the development of more general nonlinear models, particularly in fields such as machine learning, applied mathematics, and statistics. Such nonlinear models have inspired fundamental changes in which information processing systems are designed and analyzed. This emphasis on of nonlinear signal models represent a marked shift away from the linear approach pervading traditional signal processing principles. We highlight three examples of nonlinear models that have risen to the fore.

1. The *sparsity* model has emerged as a popular approach in a number of signal processing applications. Here, the fundamental assumption is that the signals of interest are sparse, i.e, that they can be represented as the sum of a few atoms selected from a fixed orthonormal basis or redundant dictionary. However, the linear sum of sparse signals is itself no longer guaranteed to be sparse, implying that the signal model is inherently nonlinear. Indeed, it can be shown that the set of sparse signals forms a highly nonlinear, union of low-dimensional subspaces residing in the ambient signal space. The sparse signal model has been successfully applied to inverse problems in a variety of application domains including medical imaging, high-speed communications, radar, and geophysics [14–17]. A particularly interesting development in this vein has been the emergence of Compressive Sensing (CS), a new paradigm for acquisition and processing of sparse signals, images, and other types of data [18–21].

2. The *low-rank* matrix model has attracted considerable interest in signal processing and machine learning. Here, the data is assumed to be well-approximated by a low-rank matrix or tensor. Once again, the set of low-rank matrices form a highly nonlinear subset of the ambient vector space of all matrices. In fact, this can be shown to the solution of a system of polynomial equations, and therefore corresponds to an algebraic variety that is a subset of the matrix space. The low-rank modeling approach has shown to be of great promise in a number of applications such as applications in collaborative filtering, multi-task learning, video analysis, and system identification [22–25].

3. The *articulation manifold* model has been the focus of significant attention, particularly in image processing. Here, the fundamental assumption is that the raw high-dimensional data is a smooth function governed by only a small number of parameters. By leveraging this concise structure, one can potentially break the "curse of dimensionality", a common problem that plagues real-world machine learning approaches. However, this functional representation is potentially highly nonlinear in its underlying parameters. Therefore, from a geometric perspective, the data is assumed to lie on a nonlinear low-dimensional submanifold of the high-dimensional space. Leveraging this nonlinear structure can greatly benefit a number of applications in machine learning, computer vision, and sensory data fusion [26–28].

1.4 Thesis Overview

This thesis is motivated by the following

> **Hypothesis**: *Next-generation signal processing systems will be manifestly nonlinear, yet extremely stable and several orders of magnitude faster than current systems.*

For this hypothesis to become a reality, there needs to be a new comprehensive framework of signal processing principles that unify, analyze, and significantly extend the scope of nonlinear models for information acquisition and processing. The ideas proposed in this thesis provide some initial steps towards building such a framework.

A particular emphasis of this thesis is the **geometry** of the signal model under consideration. In this thesis, our fundamental assumption is that the data of interest can be modeled by a *nonlinear, low-dimensional manifold* belonging to the high-dimensional, ambient signal space. We observe that a number of traditional nonlinear data models, including the three models discussed in Section 1.3, can be essentially interpreted as specific instances of such non-linear manifolds. Therefore, our proposed manifold models provide a common unifying framework that can potentially generalize to a large number of situations.

In order to understand and manipulate nonlinear manifold models, we will require a collection of handy geometric tools and techniques. We discuss a number of such tools. Paramount to our discussion are the notions of *dimension* and *smoothness* of the underlying signal manifold. We will also carefully study geometric notions such as isometry, tangent structure and geodesics in the context of high-dimensional nonlinear manifolds, and provide interpretations to particular problem instances wherever relevant.

We show that these geometric tools enable novel **algorithms** for modeling and processing high-dimensional data. Our proposed algorithms will be naturally motivated from the geometry of the signal model, and often can be viewed as appropriately constructed from smaller geometric subroutines that serve as building blocks. For example, we will frequently employ the geometric notion of a *projection* onto a manifold, which is the nonlinear generalization of the projection of a vector onto a linear subspace.

Our geometric approach also enables efficient **analysis** of our proposed models and algorithms. Wherever pertinent, we characterize the fundamental limits of the algorithms and discuss their performance in terms of their computational complexity, convergence rates, and robustness to noise and other artifacts. We also emphasize the empirical characterization of our proposed methods via a large number of numerical experiments both on simulated and real-world data. We also compare and contrast our proposed methods to existing state-of-the-art approaches based on both linear as well as nonlinear modeling principles.

Importantly, we note that our proposed geometric approach is merely one of several ways to develop nonlinear models and methods for information processing. For instance, a parallel body of research has emerged, drawing heavily from applied statistics, that focuses on *probabilistic* approaches for modeling arbitrary nonlinear functions. Instances include approaches based on *graphical models* [29–31] and *nonparametric Bayesian models* [32, 33]. However, a key distinction is that our proposed geometric approach typically enables a simpler route to rigorous analysis, as well as generalizes to a large number of problem settings of interest.

1.5 Roadmap

In this thesis, we develop new principles, algorithms, and analytical principles for nonlinear signal acquisition and processing, with a special focus on the geometry of the signal model under consideration. Our specific contributions include:

- a new convex optimization framework for designing efficient signal acquisition systems;
- a general algorithm for reconstruction, deconvolution, and denoising of signals, images, and matrix-valued data;
- efficient algorithms for inference from a small number of linear signal samples, without ever resorting to reconstruction; and,
- novel signal and image representations for robust modeling and processing of large-scale data ensembles.

We organize our specific contributions by chapter.

In **Chapter 2** we list and illustrate some fundamental mathematical principles upon which our contributions are based. We revisit some basic concepts from linear algebra, functional analysis, and differential geometry. We introduce the concept of nonlinear manifold models for signal analysis, and discuss some of their geometric properties. We study examples of signal classes described by nonlinear models, including articulation manifolds (AM), sparse and structured-sparse signals, and low-rank matrices. Additionally, we also discuss concepts in nonlinear dimensionality reduction, Compressive Sensing (CS), and affine rank minimization that will serve as the building blocks of much of the analysis and algorithms introduced in this thesis.

In **Chapter 3** we discuss the problem of designing efficient systems for the *acquisition* of signals belonging to low-dimensional nonlinear models. Suppose that the signals of interest for a given problem are well-modeled by a low-dimensional nonlinear manifold. We demonstrate that it is possible to design a signal *measurement matrix* with a minimal number of rows that preserves the geometric structure of the manifold under consideration.

Our method is based on a novel convex optimization formulation. Given a training set of points belonging to a nonlinear manifold, we pose our desired measurement matrix as a linear operator that approximately preserves the norms of all pairwise difference vectors (or *secants*) between training points. We discover this linear operator via a nuclear-norm minimization that can be efficiently solved as a convex *semi-definite program* (SDP). We develop *NuMax*, an efficient algorithm for solving our proposed SDP formulation that can be applied to massive problem sizes. We numerically demonstrate the considerable gains using the proposed NuMax algorithm over classical linear methods, such as principal components analysis (PCA).

While our primary focus in this chapter is the design of linear measurement systems for signal acquisition, our formulation can be naturally applied to related problems in machine learning, such as data retrieval using approximate nearest-neighbors (ANN), and binary classification. We demonstrate the benefits of our convex formulation in such problems.

In **Chapter 4** we consider the problem of signal *reconstruction* from noisy linear samples. We focus our attention to signals that can be written as the linear mixture of a number of components.

Our goal is to develop theoretical and practical understanding as to when it is possible to recover the component signals, given only a small number of (possibly noisy) samples of their linear sum. This is a classic problem setting and is known variously in the literature as *signal separation* or *signal deconvolution*.

More concretely, suppose that we observe noisy measurements of an unknown signal that can be modeled as the sum of two component signals, each of which arises from a nonlinear sub-manifold of a high-dimensional ambient space. We develop *Successive Projections onto INcoherent manifolds* (SPIN), a first-order projected gradient method to reconstruct the signal components. Despite the nonconvex nature of the recovery problem and the possibility of underdetermined measurements, we show that SPIN provably recovers the signal components, provided that the signal manifolds are incoherent and that the measurement operator satisfies a certain restricted isometry property.

Our formulation, and the proposed SPIN algorithm, are sufficiently general and can be extended as-is to a number of inverse problems, such as signal separation and denoising. Particularly, we demonstrate that SPIN can be applied in signal recovery applications in cases when the measurements are corrupted with structured noise (such as impulsive noise). We also apply our framework to the problem of matrix decomposition into sparse and low-rank components, and demonstrate that SPIN achieves (or even exceeds) the state-of-the-art in terms of performance.

In **Chapter 5** we consider the problem of *efficient inference* for manifold-modeled signals directly from linear signal samples. This problem setting applies to scenarios where signal recovery might only be an intermediate goal, and where the ultimate aim is to extract some salient information from the signal of interest.

We rigorously derive theoretical bounds on the performance of manifold-based inference algorithms that are only given access to a small number of random projections of signals. We prove that with the number of projections only logarithmic in the size of the original (ambient) signal space, we may (i) reliably estimate the intrinsic dimension (ID) of the underlying manifold with high accuracy, and (ii) reliably estimate the intrinsic structure of the nonlinear manifold via nonlinear dimensionality reduction techniques.

We also extend our analysis to the multi-signal scenario, where the signals are governed by a small number of common dependency parameters. We propose a new *joint manifold* model that captured such dependencies. We show that the joint manifold model can lead to improved performance for a variety of signal processing algorithms for applications including classification and manifold learning. Additionally, we leverage the random projections approach for manifolds to formulate a scalable and universal dimensionality reduction scheme that efficiently fuses the data from all sensors. Our analysis is particularly relevant in distributed sensing systems and leads to significant potential savings in data acquisition, storage and transmission costs.

In **Chapter 6** we focus our attention on the development of nonlinear models and algorithms for efficient inference on large collections of images collated "in the wild". Specifically, we introduce *novel data representations* and associated algorithms for the robust manifold modeling and processing of large-scale image ensembles. Our motivation stems from two fundamental challenges that arise while modeling real-world image collections using nonlinear low-dimensional manifolds. First, practical image manifolds are *non-isometric* to their underlying parameter space, while the state-of-the-art manifold modeling and learning frameworks assume isometry. Second, practical image manifolds are strongly perturbed by *nuisance parameters* such as illumination variations,

occlusions, and clutter.

We develop new theory and practical algorithms for manifold modeling, learning, and processing that directly address these challenges. To address the isometry challenge, we show that the Earth Movers Distance (EMD) is a more natural metric for inter-image distances than the standard Euclidean distance and use it to establish the isometry of manifolds generated by translations and rotations of a reference image. To the best of our knowledge, this is the first rigorous result on manifold isometry for generic grayscale image familes. To address the nuisance parameter challenge, we advocate an image representation based on local keypoint features and use it to define a new keypoint articulation manifold (KAM). We employ the KAM framework on a number of real-world image datasets to demonstrate its improved performance over state-of-the-art manifold modeling approaches.

Finally, in **Chapter 7** we conclude with a discussion of our contributions and highlight possible new directions for future research.

CHAPTER 2

Background on Nonlinear Signal Models

2.1 Preliminaries

2.1.1 Notation

We provide a brief review of some key mathematical principles that will underlie much of the ideas developed in this thesis. Throughout this thesis, we treat signals as real- or complex-valued functions defined over domains that are either continuous or discrete, and either infinite or finite; the various distinctions are clarified according to the context. In general, lower case boldface letters (e.g., $\mathbf{x}, \mathbf{y}$) will be used to denote discrete signals; upper case boldface letters (e.g., $\mathbf{X}, \mathbf{Y}, \mathbf{\Phi}$) will be used to denote matrices; and unbolded letters (e.g., M, N, α, β) will be used to denote scalar quantities. Calligraphic letters (e.g., $\mathcal{M}, \mathcal{X}, \mathcal{P}$) will be used to denote sets, or set-valued operators. Additional conventions will be specified as needed in each chapter.

2.1.2 Normed spaces

The modern approach to signal processing is to model signals as elements of a *normed vector space*. For example, a finite-length real signal $\mathbf{x}$ defined over a discrete domain of size N can be viewed as an element of N-dimensional space, i.e., $\mathbf{x} = [x_1, x_2, \ldots, x_N]^T \in \mathbb{R}^N$. We can endow this space with a number of different norms, including the well-known ℓ_p*-norms*,

$$\|\mathbf{x}\|_p = \begin{cases} \left(\sum_{i=1}^N |x_i|^p\right)^{\frac{1}{p}}, & p \in (0, \infty); \\ \max\limits_{i=1,2,\ldots,N} |x_i|, & p = \infty; \\ \sum_{i=1}^N \mathbf{1}_{x(i)\neq 0}, & p = 0, \end{cases} \tag{2.1}$$

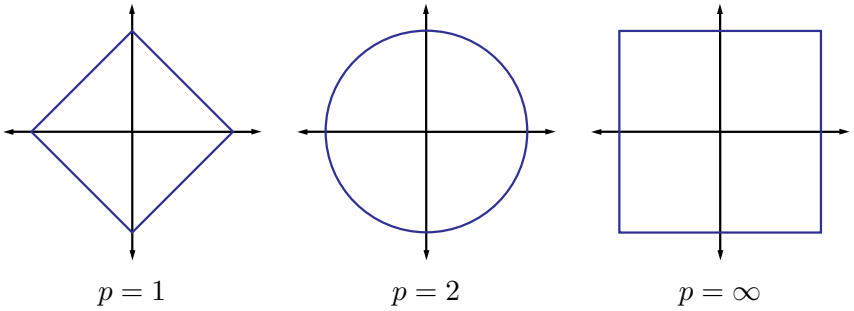


Figure 2.1: *Unit balls in* $\mathbb{R}^2$ *for the* ℓ_p *norms with* $p = 1, 2, \infty$.

where $\mathbf{1}$ denotes the indicator function.[1] A particularly common and useful norm is the ℓ_2-norm, or the *Euclidean* norm; in this case, $p = 2$. It is easy to show that the ℓ_2-norm of $\mathbf{x}$ is the norm induced by the standard inner product $\langle \cdot, \cdot \rangle$ defined on $\mathbb{R}^N$, denoted as

$$\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{y}^T \mathbf{x} = \sum_{i=1}^{N} x_i y_i. \tag{2.2}$$

In other words, $\|\mathbf{x}\|_2 = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle}$. The ℓ_2-norm is particularly useful in characterizing the *mean square error* (MSE) between pairs of signals; the MSE between $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^N$ is defined as $\frac{1}{N} \|\mathbf{x}_1 - \mathbf{x}_2\|^2$.

Analogous to the ℓ_p-norms for vectors, we can also define norms for *matrix*-valued data. We model a real-valued matrix $\mathbf{X}$ of size $M \times N$ as an element in the high-dimensional space $\mathbb{R}^{M \times N}$. We can endow this space with a number of different norms. Denote the singular value decomposition (SVD) of a matrix $\mathbf{X} \in \mathbb{R}^{M \times N}$ as $\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$, where $\mathbf{\Sigma} = \text{diag}(\boldsymbol{\sigma})$ is a diagonal, non-negative matrix where $\boldsymbol{\sigma}$ is the vector of singular values of $\mathbf{X}$. The *Frobenius norm* of $\mathbf{X}$, denoted by $\|\mathbf{X}\|_F$, is the square root of the sum of squared entries of $\mathbf{X}$, or equivalently, the ℓ_2-norm of $\boldsymbol{\sigma}$. The *nuclear norm* of $\mathbf{X}$, denoted by $\|\mathbf{X}\|_*$, is equal to the sum of its singular values, or equivalently, the ℓ_1-norm of $\boldsymbol{\sigma}$. The *spectral norm*, or the operator norm, denoted by $\|\mathbf{X}\|_1$, is the largest singular value of $\mathbf{X}$, or equivalently, the ℓ_∞-norm of $\boldsymbol{\sigma}$.

The notions of normed spaces for vectors and matrices can be elegantly extended to the case of continuous signals (giving rise to the so-called L_p*-norms*) as well as higher-order objects such as tensors (giving rise to the so-called *tensor norms*). Normed spaces serve as the focus of an important branch of functional analysis known as *Banach Theory*; see [34, 35] for details.

2.1.3 Geometry

The vector space modeling approach enable the use of concepts from *geometry*, such as lengths, distances, and angles, to describe and compare signals of interest. This insight is useful even when our signals live in high-dimensional, or even infinite-dimensional, normed vector spaces. For example, the ℓ_p-norms for signals in $\mathbb{R}^N$ have dramatically different properties for different values

[1] Note that the ℓ_p-norm does not technically satisfy the criteria for a well-defined norm when $p < 1$.

of p. This can be visualized via Fig. 2.1, in which we illustrate the unit sphere, i.e., $\{\mathbf{x} : \|\mathbf{x}\|_p = 1\}$, induced in $\mathbb{R}^2$ for $p = 1, 2, \infty$.

Consider an abstract vector space $\mathcal{F}$ representing a space of signals. We say that $\mathcal{H} \subset \mathcal{F}$ is a *linear signal model* if $\mathcal{H}$ is closed under vector addition and scalar multiplication, i.e., for any pair $\mathbf{x}, \mathbf{x}' \in \mathcal{H}$, we also have that $\alpha\mathbf{x} + \beta\mathbf{x}' \in \mathcal{H}$. However, from a geometric perspective, $\mathcal{H}$ can be viewed as a *subspace* passing through the origin of the ambient space $\mathcal{F}$. Properties of signals belonging to $\mathcal{H}$ can be understood and analyzed in terms of well-known geometric intuitions such as angles and distances along subspaces. Furthermore, suppose that a signal $\mathbf{x} \in \mathcal{H}$ is fed as input to a *linear* system (represented by the operator $\mathcal{A}$), i.e.,

$$\mathbf{y} = \mathcal{A}(\mathbf{x}). \tag{2.3}$$

Clearly, the set of all possible outputs $\mathbf{y}$ is closed under vector addition and scalar multiplication and therefore, by itself, forms a subspace $\mathcal{G} \subset \mathcal{F}$. Therefore, the system $\mathcal{A}$ can be viewed as a linear mapping between subspaces $\mathcal{A} : \mathcal{H} \mapsto \mathcal{G}$. If we are dealing with finite signals of length-N, then $\mathcal{F}$ can be viewed as N-dimensional space $\mathbb{R}^N$ and $\mathcal{A}$ can be represented in terms of a matrix $\mathbb{R}^{M \times N}$.

2.2 Manifolds

Normed vector spaces provide an elegant algebraic characterization of signals and systems that can be modeled as linear. However, in order to model more general *nonlinear* signal classes, we examine the more sophisticated mathematical framework of *low-dimensional manifolds*, commonly studied in differential geometry and topology. However, in order to model more general *nonlinear* signal classes, we examine the more sophisticated mathematical framework of *low-dimensional manifolds*, commonly studied in differential geometry and topology. For an excellent introduction to low-dimensional manifold signal models, see [36].

2.2.1 Definition

A manifold is a set that *locally* resembles a vector space around each point in the set. This emphasis on locality enables us to analyze even complex, nonlinear structures in terms of smaller building blocks that can be understood via traditional vector-space modeling techniques. The formal definition of a manifold is as follows.

Definition 1. [37, 38] *A K-dimensional manifold $\mathcal{M}$ is defined as a second-countable*[2] *Hausdorff*[3] *space that is homeomorphic*[4] *to Euclidean space $\mathbb{R}^K$.*

[2] A topological space $\mathcal{T}$ is a set X, together with a collection of subsets of X called *open sets* that satisfy the following axioms: (a) The empty set $\emptyset$ and X are open. (b) Unions of open sets are open. (c) Finite intersections of open sets are open. A topological space $\mathcal{T}$ is said to be second-countable if every open subset of $\mathcal{T}$ can be expressed as a union of elements belonging to a countable family of open sets $\mathcal{U}$.

[3] A Hausdorff space is a topological space $\mathcal{T}$ such that distinct points in $\mathcal{T}$ possess disjoint neighborhoods. In Hausdorff spaces, uniqueness of calculus concepts such as sequences, limits, and nets are well-defined.

[4] A homeomorphism is a bijective mapping between two topological spaces that is continuous and has a continuous inverse.

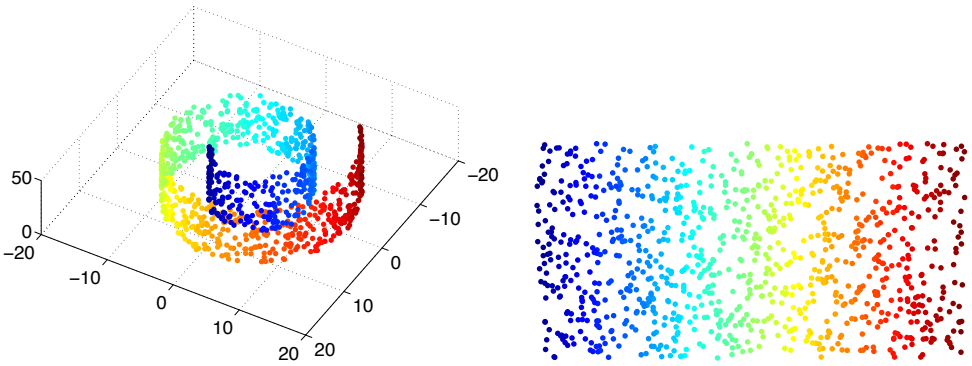


Figure 2.2: *(left) Swiss roll manifold. (right) 2D embedding of the Swiss roll.*

Loosely speaking, a manifold $\mathcal{M}$ is a set that can be "covered" by a family of smaller sets, each of which can be continuously and invertibly mapped to an open ball in Euclidean space $\mathbb{R}^K$. Each covering set is known as a *coordinate chart*, or simply, a *chart* of $\mathcal{M}$. The family of all charts is called an *atlas*. Charts are allowed to overlap, so that a single element of a manifold can be a member of different charts. While not a strict technical requirement, we will always assume that a manifold $\mathcal{M}$ is a set that is *embedded* in a higher-dimensional signal space (such as $\mathbb{R}^N$, if one considers length-N real valued signals). In other words, we may visualize $\mathcal{M}$ as a K-dimensional, nonlinear hypersurface in a larger signal space.

We list some common examples of low-dimensional manifolds:

1. Any K-dimensional subspace of Euclidean space $\mathbb{R}^N$ is a K-dimensional manifold. Here, the continuous, invertible mapping to Euclidean space is simply the identity operator defined on the subspace.

2. Any unit ℓ_p ball in $\mathbb{R}^N$ is an $(N-1)$-dimensional manifold for $N > 1$. This is easily visualized when $N = 2$ (see Fig. 2.1). For example, when $p = 2$, the unit circle in $\mathbb{R}^2$ can be broken into arcs, each of which can be continuously "unwrapped" into a line segment, i.e., a subset of $\mathbb{R}^1$.

3. Nontrivial surfaces such as the *Möbius strip* and the *Klein bottle* can also be viewed as manifolds. Despite its strange topology, a Möbius strip can be broken into smaller two-dimensional (2D) patches, each of which can be continuously mapped to a 2D subset of $\mathbb{R}^2$.

4. A common example of a manifold is the well-known *Swiss roll*, which is nothing but a 2D curved planar surface in $\mathbb{R}^3$. See Fig. 2.2 for a visual illustration of points sampled on a Swiss roll and its "unwrapped" form.

2.2.2 Geometry

We describe some important geometric properties of manifolds that prove to be useful both in theory and practice.

Tangent spaces

A K-dimensional manifold $\mathcal{M}$ is said to be *differentiable* (or *smooth*) if for any chart on $\mathcal{M}$, the function composition of the forward and inverse homeomorphisms is differentiable. Our example manifolds listed above, including the Swiss roll are all differentiable, since the local homeomorphic maps from the manifold into Euclidean space are not only continuous, but also smoothly vary from point to point.

Given a point $\mathbf{x}$ on a K-dimensional differentiable manifold $\mathcal{M} \subset \mathbb{R}^N$, we can define the notion of a *tangent space* $\mathcal{H}(\mathbf{x})$. Intuitively, $\mathcal{H}(\mathbf{x})$ can be viewed as a K-dimensional linear subspace that is formed by the span of the directional derivatives of all smooth paths on $\mathcal{M}$ that pass through $\mathbf{x}$. Any vector in $\mathcal{H}(\mathbf{x})$ is called a *tangent* to $\mathcal{M}$ at $\mathbf{x}$.

Geodesic distances

Tangent spaces enable the definition of *distances* on manifolds. Due to the subspace property, any tangent space $\mathcal{H}(\mathbf{x})$ can be equipped with an inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$. A manifold equipped with a collection of smoothly varying inner products on its tangent spaces is called a *Riemannian manifold*. The norm induced by such an inner product induces a notion of length for every tangent at $\mathbf{x}$. Therefore, the length of a smooth path on $\mathcal{M}$ can be calculated by integrating the lengths of infinitesimal tangent vectors along the path.

A natural choice of inner product is the standard inner product (2.2) in $\mathbb{R}^N$ and the corresponding norm is the ℓ_2-norm. The length of a path on a manifold using this choice of inner product is called the *geodesic distance*. Formally ,the geodesic distance between points $p, q \in \mathcal{M}$ is defined as

$$d_{\mathcal{M}}(\mathbf{p}, \mathbf{q}) = \inf\{L(\gamma) : \gamma(0) = \mathbf{p}, \gamma(1) = \mathbf{q}\}, \tag{2.4}$$

where $\gamma : [0, 1] \to \mathcal{M}$ is a C^1-smooth curve joining $\mathbf{p}$ and $\mathbf{q}$, and $L(\gamma)$ is the length of γ as measured by

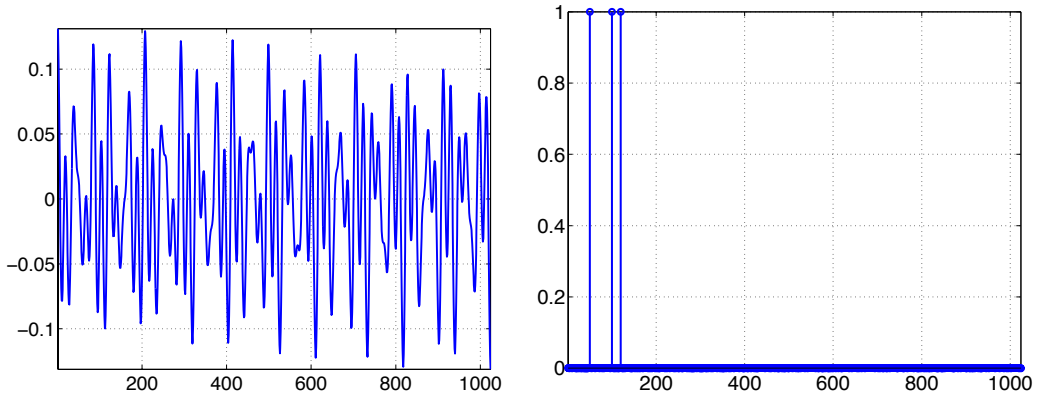
$$L(\gamma) = \int_0^1 \|\dot{\gamma}(t)\|_2 dt. \tag{2.5}$$

Condition number

It is often beneficial to study the structure of manifolds using intuitions such as "twistedness" and "self-avoidance". Such concepts are succinctly captured via the notion of *condition number*.

Definition 2. [39] *Let $\mathcal{M}$ be a Riemannian submanifold of $\mathbb{R}^N$. The* condition number *is defined as $1/\tau$, where τ is the largest number satisfying the following: the open normal bundle about $\mathcal{M}$ of radius r is embedded in $\mathbb{R}^N$ for all $r < \tau$.*

The condition number controls both local and global properties of the manifold; as $1/\tau$ becomes smaller, the manifold becomes less twisted and more self-avoiding, as observed in [39]. We will informally refer to manifolds with large τ as "good" manifolds. The condition number $1/\tau$ governs the smoothness of paths defined on $\mathcal{M}$ and can alos be used to bound the lengths of paths on $\mathcal{M}$ in terms of the geodesic distance. These properties are succinctly captured by the following Lemmata.



Signal $\mathbf{x} \in \mathbb{R}^{1024}$ DCT coefficients of $\mathbf{x}$.

Figure 2.3: *A length-1024 signal* $\mathbf{x}$ *that is 3-sparse in the Discrete Cosine Transform (DCT) basis* ($N = 1024, K = 3$).

Lemma 1. [39] *Suppose* $\mathcal{M}$ *has condition number* $1/\tau$. *Let* $p, q \in \mathcal{M}$ *be two distinct points on* $\mathcal{M}$, *and let* $\gamma(t)$ *denote a unit-speed parameterization of the geodesic path joining* p *and* q. *Then*

$$\max_t \|\ddot{\gamma}(t)\| \leq \frac{1}{\tau}.$$

Lemma 2. [39] *Suppose* $\mathcal{M}$ *has condition number* $1/\tau$. *Let* $p, q \in \mathcal{M}$ *be two points on* $\mathcal{M}$ *such that* $\|p - q\| = d$. *If* $d \leq \tau/2$, *then the geodesic distance* $d_{\mathcal{M}}(p, q)$ *is bounded by*

$$d_{\mathcal{M}}(p, q) \leq \tau(1 - \sqrt{1 - 2d/\tau}).$$

2.3 Sparse Signals

2.3.1 Definition

Given a vector space $\mathcal{F}$, a *dictionary* is simply a collection of vectors in $\mathcal{F}$ whose linear superpositions can be used to represent signals of interest. In the case of length-N signals modeled as vectors in $\mathbb{R}^N$, any dictionary can be equivalently represented as a matrix $\mathbf{D} \in \mathbb{R}^{N \times D}$, with the columns of $\mathbf{D}$ representing the dictionary vectors.

An *orthonormal basis* (ONB), or simply, a *basis* of $\mathbb{R}^N$ is a special type of dictionary that can be represented as a *square* matrix $\mathbf{\Psi}$ of size $N \times N$ whose columns are normalized and orthogonal. Given a signal $\mathbf{x} \in \mathbb{R}^N$ and a basis $\mathbf{\Psi} \in \mathbb{R}^{N \times N}$, we may *expand* $\mathbf{x}$ in terms of its basis coefficients $\boldsymbol{\alpha}$, i.e., $\mathbf{x} = \mathbf{\Psi}\boldsymbol{\alpha}$. Elementary linear algebra states that $\boldsymbol{\alpha} = \mathbf{\Psi}^{-1}\mathbf{x} = \mathbf{\Psi}^T\mathbf{x}$. Common examples of orthonormal bases relevant to signal processing include the *canonical basis* (where $\mathbf{\Psi}$ equals the identity matrix), the *Fourier basis* (where the columns of $\mathbf{\Psi}$ represent sinusoids of unit amplitude and varying frequencies), and the *wavelet basis* (where the columns of $\mathbf{\Psi}$ comprise wavelets parameterized by varying location and scale.)

Given a general dictionary $\mathbf{D}$, we are often interested in the set of signals that can be written as the linear combination of only a few elements of $\mathbf{D}$. Such signals are known as *sparse* signals. Formally, a signal $\mathbf{x} \in \mathbb{R}^N$ is said to be K-*sparse* in $\mathbf{D}$ if $\mathbf{x} = \mathbf{D}\boldsymbol{\alpha}$ and no more than K coefficients of $\boldsymbol{\alpha}$ are nonzero. The *support* of $\mathbf{x}$, $\sigma(\mathbf{x})$, is defined as the set of indices corresponding to nonzero entries of $\boldsymbol{\alpha}$. It is often convenient to represent $\sigma(\mathbf{x})$ using a binary indicator vector of length-D, with the location of the 1's indicating the indices of the nonzero coefficients. The *sparsity* of $\mathbf{x}$ is the number of nonzeros, or the ℓ_0-norm, of its dictionary representation $\boldsymbol{\alpha}$ (denoted as $\|\boldsymbol{\alpha}\|_0$).

Sparsity serves as a concise and elegant model for high-dimensional signals. A K-sparse signal $\mathbf{b} \in \mathbb{R}^N$ can be represented using merely $2K$ pieces of information, corresponding to the *locations* and *values* of the nonzero coefficients of $\boldsymbol{\alpha}$. In the regime where $K \ll N$, it is often vastly beneficial to store and process signals entirely in the space of coefficients. An example of a sparse signal in the Discrete Cosine Transform (DCT) basis is provided in Fig. 2.3; in this example, $K = 3$.

2.3.2 Geometry

It is often useful to interpret sparse signal models from a geometric perspective. Denote the set of *all* K-sparse signals as $\Sigma_K \subset \mathbb{R}^N$. One can identify Σ_K as the union of $\binom{N}{K}$ K-dimensional subspaces of $\mathbb{R}^N$, with each subspace being equivalent to the linear span of exactly K canonical unit vectors in $\mathbb{R}^N$. It is easy to show that Σ_K is a highly nonlinear subset of $\mathbb{R}^N$. For a general $\mathbf{x} \in \mathbb{R}^N$, we define its best K-sparse approximation x_K via the nonlinear optimization

$$\mathbf{x}_K = \arg \min_{\mathbf{u} \in \Sigma_K} \|\mathbf{x} - \mathbf{u}\|_2. \tag{2.6}$$

This geometric interpretation enables us to extend the notion of sparsity to more general settings. Many signals of interest exhibit more complex dependencies in terms of their nonzero values and locations. Signals that permit only a small number of admissible support configurations can be modeled by a *restricted* union of subspaces, consisting only of L_K canonical subspaces (with $L_K \ll \binom{N}{K}$). If $\Sigma = \{\sigma_1, \sigma_2, \ldots, \sigma_{L_K}\}$ denotes the restricted set of admissible supports, then a *structured sparsity model* [40] is the set

$$\mathcal{A}_K := \{\mathbf{x} : \sigma(\mathbf{x}) \in \Sigma\}. \tag{2.7}$$

2.3.3 Compressive Sensing

Compressive Sensing (CS) is a new technique for the efficient acquisition of signals, images, and other data that have a sparse representation in an basis or dictionary [18–21]. Suppose that instead of collecting all the coefficients of a vector $\mathbf{x} \in \mathbb{R}^N$, we merely record M inner products (or *measurements*) of $\mathbf{x}$ with $M < N$ pre-selected vectors. This can be represented in terms of the linear transformation $\mathbf{y} = \boldsymbol{\Phi}\mathbf{x}, \boldsymbol{\Phi} \in \mathbb{R}^{M \times N}$. The matrix $\boldsymbol{\Phi}$ is called the *sampling matrix*; it is at most rank-M and, therefore, has a nontrivial nullspace. The central result in Compressive Sensing (CS) is that, despite the non-invertible nature of $\boldsymbol{\Phi}$, if $\mathbf{x}$ is *sparse*, then it can be exactly recovered from $\mathbf{y}$ provided that $\boldsymbol{\Phi}$ satisfies a condition known as the *restricted isometry property* (RIP):

Definition 3. [20] *An $M \times N$ matrix $\boldsymbol{\Phi}$ has the K-RIP with constant δ_K if, for all $\mathbf{x} \in \Sigma_K$,*

$$(1 - \delta_K)\|\mathbf{x}\|_2^2 \leq \|\boldsymbol{\Phi}\mathbf{x}\|_2^2 \leq (1 + \delta_K)\|\mathbf{x}\|_2^2. \tag{2.8}$$

Algorithm 1 Iterative Hard Thresholding

Inputs: CS Matrix $\mathbf{\Phi}$, measurements $\mathbf{y}$, sparsity parameter K
Outputs: K-sparse approximation $\widehat{\mathbf{x}}$
Initialize: $\widehat{\mathbf{x}}_0 = 0$, $\mathbf{r} = \mathbf{y}$, $i = 0$.
while halting criterion false
  $i \leftarrow i + 1$
  $\mathbf{b} \leftarrow \widehat{\mathbf{x}}_{i-1} + \Phi^T \mathbf{r}$ {Form signal estimate}
  $\widehat{\mathbf{x}}_i \leftarrow \mathbb{M}(\mathbf{b}, K)$ {Retain best $K-$sparse approximation}
  $\mathbf{r} \leftarrow \mathbf{y} - \Phi\widehat{\mathbf{x}}_i$ {Update measurement residual}
end while
return $\widehat{\mathbf{x}} \leftarrow \widehat{\mathbf{x}}_i$

A matrix $\mathbf{\Phi}$ with the K-RIP essentially ensures a *stable embedding* of the set of *all* K-sparse signals Σ_K into a subspace of dimension M. The RIP requires $\mathbf{\Phi}$ to leave the norm of every sparse signal approximately invariant; also, $\mathbf{\Phi}$ must necessarily not contain any sparse vectors in its nullspace. At first glance, it is unclear if a matrix $\mathbf{\Phi}$ that satisfies the RIP should even exist when $M < N$; indeed, mere verification that a given sampling matrix satisfies the RIP is a very challenging problem that is widely conjectured to be NP-complete [41]. Nevertheless, it has been shown that provided $M \geq \mathcal{O}\left(K \log(N/K)\right)$, a matrix $\mathbf{\Phi}$ whose elements are i.i.d. samples from a random subgaussian[5] distribution possesses the RIP with high probability [20]. Thus, M can be linear in the sparsity of the signal set K and *only logarithmic* in the signal length N.

An analogous isometry condition holds for structured sparsity models containing L_K canonical subspaces [40, 42, 43]. This is known as the *model-based RIP* and is defined thus: $\mathbf{\Phi}$ satisfies the $\mathcal{A}_K$*-RIP* if (2.8) holds for all $\mathbf{x} \in \mathcal{A}_K$. It can be shown [42] that the number of measurements M necessary for a subgaussian sampling matrix to have the $\mathcal{A}_K$-RIP with constant δ and with probability $1 - e^{-t}$ is bounded as

$$M \geq \frac{c}{\delta^2}\left(\ln(2L_K) + K \ln \frac{12}{\delta} + t\right). \tag{2.9}$$

Note, from (2.9), that the number of measurements M is logarithmic in the *number* of subspaces in the model; thus, signals belonging to a more concise model can be sampled using fewer random linear measurements.

Given measurements $\mathbf{y} = \mathbf{\Phi x}$, CS recovery methods aim to discover the "true" sparse signal $\mathbf{x}$ that generated $\mathbf{y}$. One possible method is to seek the *sparsest* x that generates the measurements $\mathbf{y}$, i.e.,

$$\widehat{\mathbf{x}} = \arg\min_{\mathbf{x}'} \|\mathbf{x}'\|_0 \;\text{ subject to }\; \mathbf{y} = \mathbf{\Phi x}', \tag{2.10}$$

where $\|\cdot\|_0$ denotes the ℓ_0-norm. This method can be used to obtain the true solution $\mathbf{x}$, provided $M \geq 2K$ [36]. However, minimizing the ℓ_0-norm is an NP-complete problem and is not stable in the presence of noise in the measurements [20].

[5] A random variable X is called subgaussian if there exists $c > 0$ such that $\mathbb{E}(e^{Xt}) \leq e^{c^2 t^2/2}$ for all $t \in \mathbb{R}$. Examples include the Gaussian and Bernoulli random variables, as well as any bounded random variable.

Nevertheless, if the sampling matrix $\mathbf{\Phi}$ possesses the RIP, then tractable algorithms for CS recovery can be developed. These broadly follow two different approaches. The first approach entails solving a convex relaxation of (2.10):

$$\widehat{\mathbf{x}} = \arg\min_{\mathbf{x}'} \|\mathbf{x}'\|_1 \;\text{ subject to }\; \mathbf{y} = \Phi\mathbf{x}', \tag{2.11}$$

which corresponds to a linear program and hence can be solved in polynomial time. A common variant of this formulation includes accounting for noise of bounded magnitude in the measurements [44]. The second approach entails an iterative, greedy selection of the support $\sigma(\mathbf{x})$ of the true solution $\mathbf{x}$. This approach is employed by several algorithms such as orthogonal matching pursuit (OMP) [45], compressive sampling matching pursuit (CoSaMP) [46], and subspace pursuit [47]. A particularly simple greedy method, termed as iterative hard thresholding [48], is described in pseudocode form in Algorithm 1.

Both optimization and greedy approaches provide powerful stability guarantees in the presence of noise while remaining computationally efficient. Given noisy measurements of any signal $\mathbf{x} \in \mathbb{R}^N$ so that $\mathbf{y} = \mathbf{\Phi x} + \mathbf{n}$, if $\mathbf{\Phi}$ possesses the RIP, then the signal estimate $\widehat{\mathbf{x}}$ obtained by these algorithms has bounded error:

$$\|\mathbf{x} - \widehat{\mathbf{x}}\|_2 \leq C_1\|\mathbf{x} - \mathbf{x}_K\|_2 + \frac{C_2}{\sqrt{K}}\|\mathbf{x} - \mathbf{x}_K\|_1 + C_3\|\mathbf{n}\|_2, \tag{2.12}$$

where $\mathbf{x}_K$ is the best K-sparse approximation to $\mathbf{x}$ as defined in (2.6) and C_1, C_2 are global constants. Furthermore, with a simple modification, algorithms like CoSaMP and iterative hard thresholding can be used to reconstruct signals belonging to any *arbitrary* structured sparsity model [40, 49–51].

To summarize, at the core of CS lie three key concepts: a signal model exhibiting a particular type of low-dimensional geometry in high-dimensional space; a linear mapping that provides a stable embedding of the signal model into a lower dimensional space; and algorithms that perform stable, efficient inversion of this mapping.

2.4 Low-Rank Matrices

2.4.1 Definition

Sparse models provide a powerful notion of conciseness for high-dimensional signals, images, and other data. In applications pertaining to matrix-valued data, a different, sophisticated concise model can be imposed via the matrix *rank*. Consider a matrix $\mathbf{X}$ of size $M \times N$ and let $\mathbf{X} = \mathbf{U\Sigma V}^T$ be its singular value decomposition (SVD). The rank of $\mathbf{X}$ is equal to the number of nonzeros of $\boldsymbol{\sigma}$; from elementary linear algebra, we always have that $\text{rank}(\mathbf{X}) \leq \min(M, N)$. Let $\mathcal{P}_r$ denote the set of all rank-constrained matrices, i.e.,

$$\mathcal{P}_r = \{\mathbf{X} \in \mathbb{R}^{M\times N} |\;\; \text{rank}(\mathbf{X}) \leq r\}.$$

Then, $\mathcal{P}_r$ is called the rank-r *algebraic variety* of $M \times N$ matrices. The term "algebraic variety" is used since any element of $\mathcal{P}_r$ can be interpreted as the set of zeros of a system of polynomial equations [52]; specifically, any $(r+1) \times (r+1)$ minor of a matrix $\mathbf{X} \in \mathcal{P}_r$ necessarily must equal zero.

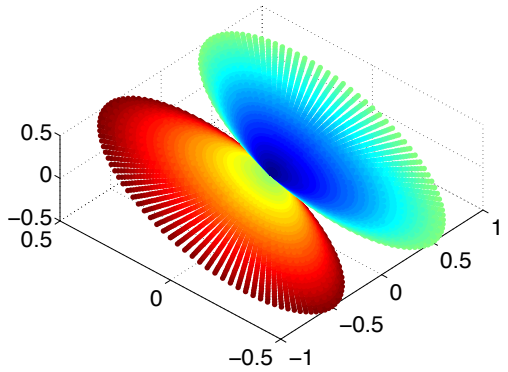


Figure 2.4: *Illustration of the manifold of symmetric rank-1 matrices embedded in* $\mathbb{R}^3$. *The set of all rank-1 matrices forms a rank-1 algebraic variety.*

2.4.2 Geometry

Analogous to the case of sparse models, we find it useful to study the low-rank matrix from a geometric perspective. It can be shown that $\mathcal{P}_r$ forms a manifold of dimension $r(M+N-r)$ that is differentiable everywhere except on a set with measure zero [53]. For cases where $r \ll \min(M, N)$, the dimension of $\mathcal{P}_r$ can be miniscule compared to the ambient dimension MN.

It is not hard to deduce that the algebraic variety $\mathcal{P}_r$ forms a highly nonlinear subset of the ambient space $\mathbb{R}^{M\times N}$. Consider the simplest case where we restrict our attention to *symmetric* matrices of size 2×2, i.e., matrices of the form

$$\mathbf{X} = \begin{bmatrix} x & y \\ y & z \end{bmatrix}.$$

Such a matrix $\mathbf{X}$ can be represented as a tuple $(x, y, z) \in \mathbb{R}^3$. Within this 3D space, the set of all matrices forms a nonlinear algebraic variety of dimension 2. We illustrate this set in Fig. 2.5.

2.4.3 Affine rank minimization

The problem of low-rank matrix recovery from a small number of linear measurements has recently attracted considerable attention from researchers in signal processing, statistics, machine learning, and control [23, 54–59]. The general technique is to pose the matrix recovery problem as the *affine rank minimization*

$$\begin{aligned} \text{minimize} \quad & \operatorname{rank}(\mathbf{X}), \\ \text{subject to} \quad & \mathcal{A}(\mathbf{X}) = \mathbf{z}, \end{aligned} \tag{2.13}$$

where the operator $\mathcal{A} : \mathbb{R}^{n\times n} \mapsto \mathbb{R}^m$ represents the linear measurements. The key insight is a convex relaxation of the minimization (2.13) to obtain the *nuclear-norm minimization*

$$\begin{aligned} \text{minimize} \quad & \|\mathbf{X}\|_*, \\ \text{subject to} \quad & \mathcal{A}(\mathbf{X}) = \mathbf{z}. \end{aligned} \tag{2.14}$$

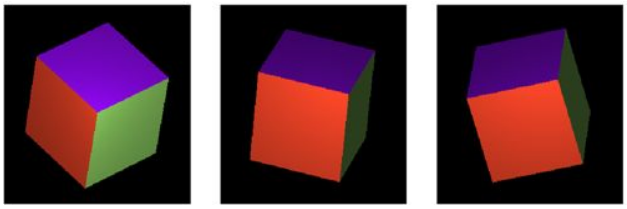


Figure 2.5: *A rotating 3D cube has 3 degrees of freedom, thus giving rise to a 3-dimensional manifold in image space.*

The nuclear-norm minimization (2.14) reduces to a semidefinite program (SDP) that admits an efficient solution in polynomial time. Fundamental analogies can be demonstrated between the problem of affine recovery of low-rank matrices and the problem of compressive sensing (CS) recovery of sparse signals. The seminal result of [23] states that the solution of the non-convex problem (2.13) exactly equals the solution of the convex problem (2.14), provided the linear map satisfies a certain restricted isometry property (RIP), exactly analogous to (2.8), but specialized to the variety of low-rank matrices. Further, if the linear measurements are of the form

$$z_i = \langle \mathbf{A}_i, \mathbf{X} \rangle \tag{2.15}$$

where the elements of $\mathbf{A}_i$ are chosen independently from a subGaussian distribution, then the corresponding operator $\mathcal{A}$ is guaranteed to satisfy the RIP with high probability. Moreover, the convex program (2.13) is exactly analogous to the ℓ_1-minimization formulation (2.11) for CS recovery.

2.5 Articulation Manifolds

2.5.1 Definition

Consider an ensemble of signals $\mathcal{M} \subset \mathbb{R}^N$ that are generated by varying K parameters $\boldsymbol{\theta} \in \Theta$, $\Theta \subset \mathbb{R}^K$. Then, we say that the signals trace out a nonlinear K-dimensional *articulation manifold* in $\mathbb{R}^N$, where $\boldsymbol{\theta}$ is called the *articulation parameter vector*. As a simple example, the set of (continuous) signals $f(t) = g(t - \theta)$, where $g(t) = e^{-t^2/2}$ is a continuous Gaussian pulse and $\theta \in \mathbb{R}$, forms a one-dimensional (1D) manifold parameterized by its mean value θ.

We focus our attention to families of high-resolution *images* governed by an articulation parameter vector $\boldsymbol{\theta}$. Examples of articulation parameters include *translation*, specifying the location of an object in a scene; *orientation*, specifying its pose; or *illumination*, specifying the positions of all light sources in a scene. Such image families form low-dimensional manifolds in the high-dimensional ambient space. We call such a family an *image articulation manifold* (IAM). The dimension K of an IAM equals the number of free parameters in the articulation $\boldsymbol{\theta}$. For example, the image translation manifold is two dimensional (2D), corresponding to horizontal and vertical translations. See Fig. 2.5 for an illustration.

Articulation manifold models for image ensembles have been extensively studied in the literature in the context of applications such as classification, data fusion, and visualization, spawning a variety of numerical algorithms [27, 60–63]. Many, if not all, of these algorithms essentially leverage

specific geometric properties of the manifold under consideration. The LLE algorithm [61], for example, works by considering local linear approximations to the manifold, which essentially the local geometric structure of the tangent spaces at different points in the IAM. We discuss some key geometric properties of IAMs below.

2.5.2 Isometry

Consider an image ensemble generated by varying a (possibly latent) articulation parameter $\boldsymbol{\theta} \in \Theta$. If Θ is a space of dimension K, then the ensemble of images forms a nonlinear image articulation manifold (IAM) $\mathcal{M} \subset \mathbb{R}^N$:

$$\mathcal{M} = \{I_{\boldsymbol{\theta}} = f(\boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}. \tag{2.16}$$

The mapping $f : \boldsymbol{\theta} \mapsto I_{\boldsymbol{\theta}}$ is said to be locally isometric if the Euclidean distance between images within a small neighborhood on the manifold $\mathcal{M}$ is proportional to the corresponding distances in the articulation space, i.e.,

$$\|I_{\boldsymbol{\theta}_1} - I_{\boldsymbol{\theta}_0}\|_2 = C\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_0\|_2, \tag{2.17}$$

for small values of $\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_0\|_2$ and a global constant C that is independent of $\boldsymbol{\theta}_0, \boldsymbol{\theta}_1$. If the local isometry property of f holds for all neighborhoods on $\mathcal{M}$, then $\mathcal{M}$ is said to be *isometric to Euclidean space*, or simply, *isometric*.

The local isometry property of IAMs has important practical ramifications. If an IAM $\mathcal{M}$ is isometric, then small variations in the articulation parameters $\boldsymbol{\theta}$ result in small variations in the image $I_{\boldsymbol{\theta}}$; in other words, the mapping f is constrained to be *stable*. More importantly, Euclidean distances measured in the space of images represent, up to a constant factor, distances measured in the parameter space $\boldsymbol{\theta}$. This enables the development of numerical methods that attempt to *infer* the latent parameters $\boldsymbol{\theta}$ given only the images $I_{\boldsymbol{\theta}}$. We revisit such methods below in Section 2.6.2.

2.5.3 Smoothness

In most practical situations, images of interest possess sharp *edges*. As images transform according to an articulation parameter $\boldsymbol{\theta}$, so too do such edges; this transformation induces an unbounded relationship between the Euclidean distance and the Euclidean distance defined on vectors in $\boldsymbol{\theta}$. More specifically, it can shown via rigorous calculations [64] that even for very simple image families, we have that

$$\|I_{\boldsymbol{\theta}_1} - I_{\boldsymbol{\theta}_2}\|_2 \geq C\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_2^{1/2}, \tag{2.18}$$

for a constant C independent of $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2$. Due to the exponent $1/2$ (instead of 1), the derivative of the function $f : \boldsymbol{\theta} \mapsto I_{\boldsymbol{\theta}}$ is unbounded everywhere. From a geometric perspective, the manifold of images containing moving edges is *nowhere differentiable*. This inherent non-smooth nature of image manifolds impedes the application of standard numerical tools from calculus and differential geometry.

The non-smooth structure of practical image articulation manifolds can be tempered using a number of techniques [64, 65]. The basic approach is to define a *smoothing functional* that acts on the individual images I; for instance, this can be a 2D Gaussian kernel ϕ_s of scale s. By applying ϕ_s to all images in the manifold $\mathcal{M}$, we obtain a new set of images that do not contain any sharp

edges; this results in a differentiable manifold $\mathcal{M}_s$ that is more amenable to numerical analysis. The parameter s can be viewed as a scale parameter; computations can be performed at a sequence of different values for s, paving the way to *multiscale* numerical methods. This is particularly useful for familiar calculus based operations, such as gradient descent and Newton's method for IAMs [65].

While the multiscale smoothing approach can render a manifold differentiable, it does not necessarily lead to isometry. Indeed, isometry is guaranteed only for manifolds of black-and-white images exhibiting certain types of restrictive symmetries [64]. However, for a pair of generic grayscale images $I_{\boldsymbol{\theta}_1}, I_{\boldsymbol{\theta}_2}$ belonging to $\mathcal{M}$, the distance metric $d_{\mathcal{M}}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ computed between the images smoothed at scale s is not necessarily proportional to $\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_2$ for any choice of the parameter s. In Chapter 6 we develop a computational framework that directly mitigates the challenges posed by the non-isometry of real-world image ensembles.

2.6 Dimensionality Reduction

2.6.1 Principal Components Analysis

Consider an ensemble of signals represented by Q vectors $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_Q\} \subset \mathbb{R}^N$, where both N, Q are potentially very large. Suppose we represent the elements of $\mathcal{X}$ as columns of a matrix $\mathbf{X} \in \mathbb{R}^{N \times Q}$, which we term as the *data matrix*. Given such a high-dimensional dataset, a natural question is whether it can be represented by (or *embedded into*) a lower-dimensional space with minimal distortion. Such an embedding would enable tractable storage and processing of the dataset without sacrificing performance accuracy.

One such embedding can be obtained via a popular technique in statistics known as principal components analysis (PCA), sometimes referred to as the Karhunen-Loève Transform [66]. PCA proceeds as follows; given a data matrix $\mathbf{X}$, we perform an SVD of $\mathbf{X}$, i.e., $\mathbf{X} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T$. Next, we *linearly* project the columns of $\mathbf{X}$ onto the subspace spanned by the r leftmost columns of $\mathbf{U}$, denoted by $\mathbf{U}_r$. The projected columns are called the *principal components* of the data $\mathbf{X}$. Suppose we represent the projected columns of $\mathbf{X}$ as a rank-r matrix $\mathbf{X}_r = \mathbf{U}_r\boldsymbol{\Sigma}_r\mathbf{V}_r^T$ of size $N \times Q$. It is well known that $\mathbf{X}_r$ is the optimal approximation to $\mathbf{X}$ in terms of the Frobenius norm, i.e.,

$$\|\mathbf{X} - \mathbf{X}_r\|_F \leq \|\mathbf{X} - \mathbf{Y}\|_F ,$$

where $\mathbf{Y}$ is any other rank-r matrix. Therefore, PCA can be viewed as an efficient *linear* embedding method that minimally distorts the data in terms of the Frobenius norm. PCA has witnessed widespread use in machine learning and statistics, and has successfully been applied for signal acquisition, compression, estimation, and classification [67, 68]. Further, PCA can be adapted to account for problem-specific requirements. For example, if the data vectors originate from one of two classes and the goal is to maintain class separability, then PCA can be modified to produce new, related linear embedding techniques such as Linear Discriminant Analysis (LDA) and Factor Analysis [69, 70].

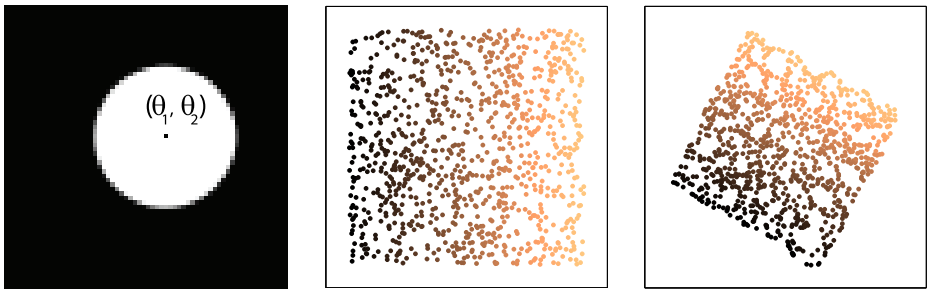


Figure 2.6: *(a) Input data consisting of 1000 images of a disk shifted in $K = 2$ dimensions, parametrized by an articulation vector $(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$. (b) True $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ values of the sampled data. (c) Isomap embedding learned from original data in $\mathbb{R}^N$. The Isomap embedding*

2.6.2 Manifold learning

Principal Components Analysis (PCA) is particularly well suited for scenarios where the data originates from a *subspace* of the high-dimensional ambient space. On the other hand, data modeled by highly nonlinear manifolds cannot be well approximated by linear subspaces. This has led to the development of *nonlinear* algorithms for dimensionality reduction. Such algorithms produce a nonlinear mapping that embeds the data in a low-dimensional space, simultaneously trying to minimize a suitable objective function that represents the fidelity of the mapping in a local, or global, sense.

The nonlinear embedding of data into a low-dimensional space is sometimes referred to as *manifold learning*. Manifold learning techniques are particularly useful in inferring the latent parameters of signals originating from articulation manifolds. Consider a K-dimensional image articulation manifold $\mathcal{M} = \{I_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$. Suppose that we are given access to a dataset of n images $I_{\boldsymbol{\theta}_1}, I_{\boldsymbol{\theta}_1}, \ldots, I_{\boldsymbol{\theta}_n}\} \subset \mathcal{M}$, where the articulation parameters are unknown. Then, a nonlinear embedding into a K-dimensional space, obtained by running a manifold learning algorithm on the input dataset, often reveals a faithful approximation to the corresponding unknown parameters $\{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_n\}$.

A host of algorithms for manifold learning have been developed, and some well-known examples include Locally Linear Embedding [61], Laplacian Eigenmaps [71], Hessian Eigenmaps [72], Maximum Variance Unfolding [73], and Diffusion Maps [74]. We focus on Isomap [60], a popular manifold learning algorithm. An illustration of the performance of Isomap is provided in Fig. 2.6. We consider a dataset of $n = 1000$ images of shifts of a white disc on a black background. Here, the underlying IAM is parameterized by the 2D location of the center of the disc $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$. The true locations of the disc centers are recorded and the 2D embeddings of the images are obtained using Isomap. Figure 2.6 demonstrates that the 2D embedding of the images closely approximates the true parameters (modulo a 2D planar rotation). Several ideas proposed and developed in Chapters 5 and 6 heavily rely upon this basic intuition.

Isomap works in three stages:

1. Construct a graph G that contains a vertex for each data point; an edge connects two vertices if the Euclidean distance between the corresponding data points is below a specified threshold.

Algorithm 2 Isomap

Inputs: Data $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\} \subset \mathbb{R}^N$, embedding dimension K.
Output: Low-dimensional data representation $\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n\} \subset \mathbb{R}^K$

1. Construct connectivity graph $G = (V, E)$
2. Calculate distances $\widetilde{D}_{ij} = \|\mathbf{x}_i - \mathbf{x}_j\|_2^2$ for $(i, j) \in E$
3. Calculate distances $\widetilde{D}_{ij}$ for $(i, j) \notin E$ using a shortest-path algorithm.
4. Center to make zero mean: $\widetilde{\mathbf{D}} \leftarrow -\mathbf{C}\widetilde{\mathbf{D}}\mathbf{C}, \quad \mathbf{C} = \mathbf{I} - \mathbf{1}\mathbf{1}^T/n$
5. Perform eigen-decomposition $\widetilde{\mathbf{D}} = \mathbf{U}\Sigma\mathbf{U}^T$
6. Return top K eigenvectors $\mathbf{Y} = \Sigma_K^{1/2}\mathbf{U}_K$.

2. Weight each edge in the graph G by computing the Euclidean distance between the corresponding data points. Then, estimate the geodesic distance between each pair of vertices as the length of the shortest path between the corresponding vertices in the graph G.

3. Embed the points in $\mathbb{R}^K$ using multidimensional scaling (MDS) [75].

The Isomap algorithm is described in pseudocode form in Algorithm 2. A crucial component of the MDS step is a suitable linear transformation of the matrix of squared geodesic distances; the rank-r approximation of this new matrix yields the best possible r-dimensional coordinate structure of the input sample points in a mean-squared sense. Further results on the performance of Isomap in terms of geometric properties of the underlying manifold can be found in [76].

The minimum squared error in MDS is called *stress*; in the manifold learning literature, it is usually referred to as *residual variance*. The residual variance R serves as a global metric for how well Isomap manages to embed the input data; smaller values of R imply better embeddings, and vice versa. In Chapter 5 we introduce efficient methods to calculate Isomap embeddings for very high-dimensional datasets, and characterize our methods via a careful residual variance analysis.

CHAPTER 3

Near-Isometric Linear Embeddings of Manifolds

3.1 Setup

In this chapter, we focus on the problem of low-dimensional representations, or *embeddings* of data that can be modeled as elements of a high-dimensional space. This problem assumes relevance for a number of tasks in the information processing pipeline, including data acquisition, compression, and inference. As introduced in Section 2.6, a classical dimensionality reduction technique is principal components analysis (PCA) [66], which involves constructing a *linear* mapping the original N-dimensional data into the K-dimensional subspace spanned by the dominant eigenvectors of the data covariance matrix; typically, $K \ll N$. Over the last decade, more sophisticated *nonlinear* data embedding methods, known as manifold learning algorithms, have also emerged [60, 61, 72, 79].

Linear dimensionality reduction is advantageous in several aspects. A key appeal of linear dimensionality reduction methods is their *computational efficiency*. Techniques such as PCA can be very efficiently performed using a singular value decomposition (SVD) of a linear transformation of the data. Another key appeal is their *generalizability*: linear methods produce smooth, globally defined mappings that can be easily applied to unseen, out-of-sample test data points. Nevertheless, PCA, and its variants, are marked by the important shortcoming that the generated embedding potentially *distorts pairwise distances* between sample data points. This phenomenon is exacerbated when the data arises from a nonlinear submanifold of the signal space [80]. Due to this behaviour, two distinct points in the ambient signal space are often be mapped to a single point in the low-dimensional embedding space. This hampers the application of PCA-like techniques to some important problems such as reconstruction and parameter estimation of manifold-modeled signals.

An intriguing alternative to PCA is the method of *random projections*. Consider $\mathcal{X}$, a cloud of

This work is in collaboration with Richard G. Baraniuk, Aswin C. Sankaranarayanan, and Wotao Yin [77, 78].

Q points in a high-dimensional Euclidean space $\mathbb{R}^N$. The Johnson-Lindenstrauss Lemma [81] states that $\mathcal{X}$ can be linearly mapped to a subspace of dimension $M = \mathcal{O}(\log Q)$ with minimal distortion of the pairwise distances between the Q points (in other words, the mapping is *near-isometric*). Further, this linear mapping can be easily implemented in practice. One simply constructs a matrix $\mathbf{\Phi} \in \mathbb{R}^{M \times N}$ with $M \ll N$ whose elements are randomly drawn from certain probability distributions. Then, with high probability, $\mathbf{\Phi}$ is approximately isometric under a certain lower-bound on M [80, 81]. The method of random projections can be extended to more general signal classes beyond finite point clouds. For example, random linear projections provably preserve the isometric structure of compact, differentiable low-dimensional manifolds [82, 83], as well as the isometric structure of the set of sparse signals [20, 84]. Random projections are conceptually simple and useful in many signal processing applications. Yet, they too suffer from certain shortcomings. Their theoretical guarantees are probabilistic and asymptotic. Further, the mapping itself is independent of the data and/or task under consideration and hence cannot leverage any special geometric structure of the data if present.

In this chapter, we propose a novel framework for *deterministic* construction of *linear, near-isometric* embeddings of a finite set of data points. Given a set of training points $\mathcal{X} \subset \mathbb{R}^N$, we consider the *secant set* $\mathcal{S}(\mathcal{X})$ that consists of all pairwise difference vectors of $\mathcal{X}$, normalized to lie on the unit sphere. Next, we formulate an affine-rank minimization problem (3.3), subject to max-norm constraints, to construct a matrix Ψ that preserves the norms of all the vectors in $\mathcal{S}(\mathcal{X})$ up to a desired distortion parameter δ. Affine rank-minimization is known to be NP-hard and so we perform a convex relaxation to obtain a trace-norm minimization (3.4), which is equivalent to a tractable semi-definite program (SDP).

The semi-definite program (3.4) can be solved using any generic interior-point method for convex programming (for example, CVX [85]). However, the convergence of such solvers, although guaranteed, is typically very slow even for small problem sizes. Further, the presence of the max-norm constraints (3.4), though convex, negates the direct application of existing first-order methods for large-scale semidefinite programming [86]. In order to enable scalability of our method to large-scale problems, we adopt a two-stage approach. First, we develop a novel algorithm that we call *Nuclear norm minimization with Max-norm constraints* (NuMax) to solve (3.4). Our proposed NuMax algorithm is based on the Alternating Direction Method of Multipliers (ADMM) and exhibits much faster rates of convergence than standard approaches. Second, we develop a greedy approximate method to solve the SDP (3.4) based on the *Column Generation* approach commonly used to solve large-scale linear programs. This enables us to apply our framework even to problems where the total number of elements in the secant set $\mathcal{S}(\mathcal{X})$, i.e., the number of constraints in (3.4), grows intractably large and is potentially greater than the memory available for computation.

We show that our optimization framework for designing linear embeddings is useful for a number of applications in machine learning and signal processing. First, if the training set $\mathcal{X}$ comprises sufficiently many points that are uniformly drawn from a low-dimensional signal manifold $\mathcal{M}$, then the calculated matrix Ψ represents an near-isometric linear embedding over all pairwise secants of $\mathcal{M}$. In other words, Ψ satisfies the restricted isometry property (RIP) for signals belonging $\mathcal{X}$, and therefore our method enables the design of *efficient sensing matrices* for Compressive Sensing (CS) applications. Second, by carefully pruning the secant set $\mathcal{S}(\mathcal{X})$, we can tailor our proposed linear embedding to enable more general signal inference tasks, such as *supervised binary classification.*

Several numerical experiments in Section 3.6 demonstrate the advantages of our approach.

This chapter is organized as follows. In Section 3.2 we provide a brief background on existing methods for linear dimensionality reduction. In Section 3.3 we introduce our main theoretical contributions and propose an SDP formulation for designing "good" linear embeddings. In Section 3.5 we develop computationally efficient algorithms that can solve our proposed SDP for large-scale problems. In Section 3.6 we apply our linear embedding framework to a number of diverse problems and demonstrate its efficiency both on synthetic and real-world datasets. In Section 3.7 we provide some concluding remarks.

3.2 Related Work

3.2.1 Linear dimensionality reduction

The problem of constructing low-dimensional geometry-preserving embeddings (i.e., embeddings that retain all, or some part of, the distance information between data points) is a well-studied problem in machine learning, applied statistics, and computational geometry. For an excellent overview of this topic, see [87].

The classical approach for dimensionality reduction is Principal Components Analysis (PCA), introduced in Section 2.6. PCA has witnessed widespread use in machine learning and statistics, and has successfully been applied for signal acquisition, compression, estimation, and classification [67, 68]. PCA is conceptually simple and efficient. The only computational requirements involve performing the SVD of the data matrix to find the leading singular vectors, and projecting the data onto a basis for the subspace spanned by these singular vectors.

Additionally, owing to its linear nature, PCA can easily be extended to *out-of-sample* points. Given a training dataset $\mathbf{X}$, a low-dimensional embedding of a new, unseen data vector $\mathbf{x}' \notin \mathbf{X}$ simply involves a linear projection of $\mathbf{x}'$ on the PCA basis vectors. Further, PCA can be adapted to account for problem-specific requirements. For example, if the data vectors originate from one of two classes and the goal is to maintain class separability, then PCA can be modified to produce new, related linear embedding techniques such as Fisher's Linear Discriminant Analysis (LDA) and Factor Analysis [69, 70].

Nevertheless, PCA is accompanied by certain drawbacks. Crucially, the optimality of PCA is not accompanied by any guarantees regarding local geometric properties of the resulting embedding [80]. Therefore, any information contained in geometric inter-relationships between data points is irrevocably lost. One can conjure examples of datasets where the distance between the PCA embeddings of two distinct high-dimensional points is vanishingly small. Thus, PCA embeddings are not guaranteed to be *isometric* (i.e., distance-preserving), or even *invertible*. This profoundly affects both algorithmic design and analysis, and fundamental problems, such as reconstruction of signals from their PCA embeddings, cannot be satisfactorily addressed.

One way to resolve the issue is to relax the isometry requirement to some extent. Instead of exactly trying to preserve all pairwise distances between data points, a desirable embedding might only approximately preserve pairwise distances. The following definition makes this intuition concrete.

Definition 4. *Let $\mathcal{X} \subset \mathbb{R}^N$ be a finite set of points. Suppose $M \leq N$. An embedding operator $\mathcal{P} : \mathbb{R}^N \to \mathbb{R}^M$ satisfies the* restricted isometry property (RIP) *on $\mathcal{X}$ if there exists a positive constant $\delta > 0$ such that for every $\mathbf{x}, \mathbf{x}'$ in $\mathcal{X}$, the following relations hold:*

$$(1-\delta)\|\mathbf{x}-\mathbf{x}'\|_2^2 \leq \|\mathcal{P}\mathbf{x}-\mathcal{P}\mathbf{x}'\|_2^2 \leq (1+\delta)\|\mathbf{x}-\mathbf{x}'\|_2^2. \tag{3.1}$$

Notice that (3.1) is identical to (2.8), except that the relaxed notion of isometry is defined over a finite point set. The quantity δ encapsulates the deviation from perfect isometry, and is called the *isometry constant*. We (trivially) observe that the identity operator on $\mathcal{X}$ always satisfies the RIP with $\delta = 0$; however, in this case $M = N$. The main question, from a dimensionality reduction perspective, is whether embedding operators that satisfy the RIP exist for $M < N$. The celebrated *Johnson-Lindenstrauss* (JL) Lemma answers this in the affirmative [81].

Lemma 3 (JL)**.** *Consider a dataset $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_P\} \subset \mathbb{R}^N$. Let $M \geq \mathcal{O}\left(\delta^{-2} \log P\right)$. Construct a matrix $\mathbf{\Phi} \in \mathbb{R}^{M \times N}$ such that each element of $\mathbf{\Phi}$ is drawn independently from a Gaussian distribution with zero mean and variance $1/M$. Then, with high probability, the linear operator $\mathbf{\Phi} : \mathbb{R}^N \to \mathbb{R}^M$ satisfies the RIP on $\mathcal{X}$.*

Therefore, in order to construct a (near) isometric embedding of a dataset $\mathcal{X}$, one simply constructs a matrix $\mathbf{\Phi}$ by choosing elements randomly from a Gaussian distribution. Other kinds of distributions are also admissible [88, 89].

We highlight some important features of the JL Lemma. First, like PCA, the constructed embedding $\mathbf{\Phi}$ is *linear*; therefore it is computationally efficient and can be applied to out-of-sample points. Second, unlike PCA, the constructed embedding $\mathbf{\Phi}$ is *universal*, i.e., fully independent of the dataset $\mathcal{X}$. Instead of projecting the data onto the basis vectors of the subspace formed by the singular vectors of $\mathcal{X}$, one simply picks a few basis vectors at random, and projects the data onto these vectors. The JL Lemma guarantees that such an embedding also preserves the local geometric structure of $\mathcal{X}$. Third, the dimension of the lower-dimensional embedding M is only logarithmic in the number of data points, and is independent of the ambient dimension N; therefore, potentially $M \ll N$.

Thus, the random projections approach provides a simple, universal method to construct linear embeddings that satisfy the RIP for arbitrary datasets. In fact, it can be shown that in the worst case, for a given isometry constant δ, it is not possible to embed the dataset $\mathcal{X}$ into any M-dimensional space where $M \leq \delta^{-2} \log(\delta^{-1}) \log P$ [90]. However, this worst case is pathological and rarely occurs in practice. Also, the universality of random projections negates the ability to construct embeddings that leverage the intrinsic geometry of a given set of data vectors. Therefore, we pose the following natural question: for a specific dataset $\mathcal{X}$ and a parameter δ, what is the dimension of the smallest possible space into which $\mathcal{X}$ can be embedded with distortion δ? Our proposed framework in Section 3.3 takes some initial steps towards answering this question.

3.2.2 Secant-preserving embeddings

Our motivation for constructing efficient near-isometric linear embeddings stems from a result in differential geometry known as *Whitney's Embedding Theorem* [37]. This result states that any

smooth K-dimensional submanifold $\mathcal{M}$ can be mapped down to a Euclidean space of dimension $2K+1$ such that the isometric structure of $\mathcal{M}$ is preserved. An important notion in the proof of Whitney's Theorem is the normalized *secant manifold* of $\mathcal{M}$:

$$\mathcal{S}(\mathcal{M}) = \left\{ \frac{\mathbf{x}-\mathbf{x}'}{\|\mathbf{x}-\mathbf{x}'\|_2}, \quad \mathbf{x},\mathbf{x}' \in \mathcal{M}, \mathbf{x} \neq \mathbf{x}' \right\}. \tag{3.2}$$

The secant manifold $\mathcal{S}(\mathcal{M})$ forms a $2K$-dimensional submanifold of the $(N-1)$-dimensional unit sphere in $\mathbb{R}^N$. Any point on the unit sphere represents a projection from $\mathbb{R}^N$ to $\mathbb{R}^{N-1}$. Therefore, by choosing a projection direction that does *not* belong to $\mathcal{S}(\mathcal{M})$, one can map $\mathcal{M}$ into $\mathbb{R}^{N-1}$ injectively (i.e., without overlap). If $2K \leq N-1$, then this is always possible. Whitney's (weak) Embedding Theorem is based upon this intuition. Unfortunately, the proof for Whitney's Theorem is *non-constructive* and therefore does not lend itself to easy implementation in practice.

In Section 3.3 below, we develop an efficient computational framework to produce a low-dimensional linear embedding that satisfies the RIP (3.1) on a given dataset $\mathcal{X}$. Specifically, we build upon and improve the *Whitney Reduction Network* (WRN) approach for computing auto-associative graphs [91, 92]. The WRN approach is a heuristic technique that is algorithmically similar to PCA. Our approach follows a completely different path. The optimization formulation (3.4) is based on convex programming and is guaranteed to produce a near-isometric, linear embedding. Further, in Section 3.5, we develop efficient algorithms to solve the optimization (3.4) for large-scale problems.

3.3 A Convex Approach for Designing Isometric Embeddings

3.3.1 General framework

Given a dataset $\mathcal{X} \subset \mathbb{R}^N$, our goal is to find a linear embedding $\mathcal{P} : \mathbb{R}^N \to \mathbb{R}^M$, $M \ll N$, that satisfies the RIP (3.1) on $\mathcal{X}$. We form the secant set $\mathcal{S}(\mathcal{X})$ using (3.2) to obtain a set of $Q' = \binom{Q}{2}$ unit vectors $\mathcal{S}(\mathcal{X}) = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_{Q'}\}$. We seek a *measurement matrix* $\mathbf{\Psi} \in \mathbb{R}^{M \times N}$ with as few rows as possible that satisfies the RIP on $\mathcal{S}(\mathcal{X})$ with an isometry constant δ.

We cast the above problem in terms of a feasible optimization over the space of PSD symmetric matrices. Let $\mathbb{S}^{N \times N}$ be the set of symmetric $N \times N$ matrices. Define $\mathbf{P} \doteq \mathbf{\Psi}^T\mathbf{\Psi} \in \mathbb{S}^{N \times N}$; then, we have that $\text{rank}(\mathbf{P}) = M$. Also, we have the constraints that $|\,\|\mathbf{\Psi}\mathbf{v}_i\|_2^2 - 1| = \left|\mathbf{v}_i^T\mathbf{P}\mathbf{v}_i - 1\right|$ is no greater than δ for every secant $\mathbf{v}_i$ in $\mathcal{S}(\mathcal{X})$. Suppose that b represents the all-ones vector $\mathbf{1}_{Q'}$, and that $\mathcal{A}$ represents the linear operator that maps a symmetric matrix $\mathbf{X}$ to the Q'-dimensional vector $\mathcal{A} : \mathbf{X} \to \{\mathbf{v}_i^T\mathbf{X}\mathbf{v}_i\}_{i=1}^{Q'}$. Then, we can formulate $\mathbf{P}$ as the solution to the optimization problem

$$\begin{aligned} \text{minimize} \quad & \text{rank}(\mathbf{P}) \\ \text{subject to} \quad & \mathbf{P} \succeq 0, \ \mathbf{P} = \mathbf{P}^T, \\ & \|\mathcal{A}(\mathbf{P}) - \mathbf{b}\|_\infty \leq \delta. \end{aligned} \tag{3.3}$$

Affine rank minimization is a highly nonconvex problem and is known to be NP-hard in general [23]. Therefore, following the approach in [93], we instead propose solving a nuclear-norm

relaxation of (3.3):

$$\begin{aligned} \text{minimize} \quad & \|\mathbf{P}\|_* \\ \text{subject to} \quad & \mathbf{P} \succeq 0,\ \mathbf{P} = \mathbf{P}^T, \\ & \|\mathcal{A}(\mathbf{P}) - \mathbf{b}\|_\infty \leq \delta. \end{aligned} \tag{3.4}$$

Since $\mathbf{P}$ is a PSD symmetric matrix, the nuclear norm of $\mathbf{P}$ is equal to its trace. Thus, the matrix recovery problem (3.4) consists of minimizing a linear objective function subject to linear inequality constraints over the cone of PSD symmetric matrices, and hence is equivalent to a semi-definite program (SDP), which can be solved in polynomial time [94]. Once the solution $\mathbf{P}^* = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T$ to (3.4) is found, the desired linear embedding $\mathbf{\Psi}$ can be calculated using a simple matrix square root:

$$\mathbf{\Psi} = \mathbf{\Lambda}_M^{1/2}\mathbf{U}_M^T, \tag{3.5}$$

where $\mathbf{\Lambda}_M = \text{diag}\{\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_M\}$ denotes the M leading (non-zero) eigenvalues of $\mathbf{P}^*$. In this manner, we obtain a low-rank matrix $\mathbf{\Psi} \in \mathbb{R}^{M \times N}$ that satisfies the RIP on the secant set $\mathcal{S}(\mathcal{X})$ with isometry constant δ. The convex optimization formulation (3.4) is conceptually very simple, and the only inputs to (3.4) is the input dataset $\mathcal{X}$ and the desired isometry constant $\delta > 0$.

3.3.2 Analysis

Since we seek a measurement matrix $\mathbf{\Psi}$ with a minimal number of rows, a natural question to ask is whether the nuclear-norm relaxation (3.4) is guaranteed to produce solutions $\mathbf{P}^*$ of minimum rank. Indeed, the efficiency of nuclear-norm minimization for low-rank matrix recovery has been thoroughly examined [23, 95, 96]. However, the best known theoretical results for (3.4) make certain restrictive assumptions on the linear operator $\mathcal{A}$; for instance, one common assumption is that the entries of the matrix representation of $\mathcal{A}$ are independently drawn from a standard normal distribution. This assumption is clearly violated in our case, since $\mathcal{A}$ is a function of the secant set $\mathcal{S}(\mathcal{X})$ and in general heavily depends on the geometry of the data at hand. Nevertheless, the following classical result in semidefinite programming provides the following upper bound of the rank of the optimum $\mathbf{P}^*$.

Proposition 1 ([97, 98])**.** *Let r^* be the rank of the optimum to the semi-definite program (3.4). Then,*

$$r^* \leq \left\lceil \frac{\sqrt{8|\mathcal{S}(\mathcal{X})| + 1} - 1}{2} \right\rceil$$

The upper bound on the optimal rank r^* provided in Proposition 1 is loose since the cardinality of $\mathcal{S}(\mathcal{X})$ can potentially be very large. Additionally, one might intuitively expect the optimal rank r^* to depend upon the geometric arrangement of the data vectors in $\mathcal{X}$, as well as the input isometry constant δ; however, the bound in Proposition 1 does not reflect this dependence. A full analytical characterization of the optimal rank obtained by the program (3.4) is of considerable interest both in theory and practice, but we will not pursue that direction here.

3.4 Extensions

3.4.1 Linear embeddings of manifolds

The optimization framework (3.3) provides a novel method for producing low-dimensional embeddings of data that can be modeled as submanifolds of the signal space. The use of semidefinite programming approaches for manifold-modeled data has been addressed in the literature, but the focus has been on producing general nonlinear embeddings of point clouds [79, 99]. Due to its special structure, the optimization formulation (3.3) produces an explicit *linear* embedding operator $\mathcal{P}$, equivalent to the matrix $\mathbf{\Psi} \in \mathbb{R}^{M \times N}$.

Furthermore, if the training data $\mathcal{X}$ originates from a K-dimensional smooth manifold $\mathcal{M} \subset \mathbb{R}^N$, then the main result in [82] establishes that the near-isometric property of the proposed embedding $\mathcal{P}$ on pairwise secants of X extends to *all* pairwise secants in $\mathcal{M}$. More precisely, suppose that $\mathbf{\Psi}$ satisfies the RIP (3.1) with constant δ over a training set $\mathcal{X} \in \mathbb{R}^N$. If the training set $\mathcal{X}$ represents a high-resolution sampling of points on or close to $\mathcal{M}$, then $\mathbf{\Psi}$ provably satisfies the RIP with a slightly larger constant over the *entire manifold* $\mathcal{M}$. See Section 3.2.5 of [82] for a detailed discussion of this approach.

We numerically demonstrate this phenomenon on synthetic manifold-modeled data below in Section 3.6. An important challenge in such a context is the choice of training data $\mathcal{X}$. The proof techniques of [82] assume that the training set $\mathcal{X}$ is an ϵ-cover of $\mathcal{M}$, i.e., for every $\mathbf{m} \in \mathcal{M}$, there exists an $\mathbf{x} \in \mathcal{X}$ such that $\min_{\mathbf{x} \in \mathcal{X}} d_{\mathcal{M}}(\mathbf{m}, \mathbf{x}) \leq \epsilon$ for a small constant $\epsilon > 0$. However, covering results from high-dimensional geometry state that the cardinality of such a set $\mathcal{X}$, in the worst case, can be exponential in the manifold dimension K, i.e., $|\mathcal{X}| = \mathcal{O}\left(\left(\frac{1}{\delta}\right)^K\right)$. In real-world problems, computations involving training sets of such large training sets $\mathcal{X}$ may be intractable, and thus one may have to resort to heuristic, computationally efficient methods in practice.

3.4.2 Class-specific linear embeddings

We observe that the matrix inequality constraints in (3.4) are derived by enforcing an approximate isometry condition on *all* pairwise secants $\{\mathbf{v}_i\}_{i=1}^{Q'}$. However, this can often prove to be too restrictive. For example, consider a supervised classification scenario where the signal of interest $\mathbf{x}$ can arise from one of two classes that are each modeled by low-dimensional manifolds $\mathcal{M}, \mathcal{M}'$. The goal is to infer the signal class of $\mathbf{x}$ from the linear embedding $\mathcal{P}\mathbf{x}$. Here, the measurement matrix $\mathbf{\Psi}$ should be designed such that signals from different classes are mapped to sufficiently distant points in the low-dimensional space. However, it does not matter if signals from the same class are mapped to the same point.

Geometrically speaking, we seek $\mathcal{P}$ that preserves merely the *inter-class secants*

$$\mathcal{S}(\mathcal{M}, \mathcal{M}') = \left\{ \frac{\mathbf{x} - \mathbf{x}'}{\|\mathbf{x} - \mathbf{x}'\|_2}, \;\; \mathbf{x} \in \mathcal{M}, \; \mathbf{x}' \in \mathcal{M} \right\} \tag{3.6}$$

up to an isometry constant δ. This represents a reduced set of linear constraints in (3.4). Therefore, the solution space to (3.4) is larger, and the optimal embedding $\mathcal{P}^*$ will necessarily be of lower rank, thus leading to a further reduction in the dimension of the embedding space. We examine the problem of finding class-specific linear embeddings in greater detail below in Section 3.6.

3.5 Efficient Algorithms

The semidefinite program (3.4) admits a tractable solution in polynomial time using interior-point methods. However, for a generic SDP with Q' constraints and a matrix variable of size $N \times N$, interior-point methods incur memory costs that scale as $\mathcal{O}\left(Q'^2\right)$ and time-complexity costs that scale as $\mathcal{O}\left(N^6\right)$. Therefore, solving (3.4) using traditional semidefinite programming packages [100, 101] becomes infeasible for practical real-world applications. Below, we develop two algorithms that exploit the special structure of the optimization problem (3.4) and produce very efficient solutions at vastly reduced costs.

3.5.1 An ADMM approach

We develop an efficient algorithm to solve (3.4) based on the Alternating Direction Method of Multipliers (ADMM). We call our algorithm *NuMax*, an abbreviation for *Nuclear norm minimization with Max-norm constraints*. We rewrite (3.4) by introducing auxiliary variables $\mathbf{L} \in \mathbb{S}^{N\times N}$ and $\mathbf{q} \in \mathbb{R}^{Q'}$. We obtain the optimization problem

$$\begin{aligned} \text{minimize} \quad & \|\mathbf{P}\|_* && (3.7)\\ \text{subject to} \quad & \mathbf{P} = \mathbf{L}, \quad \mathcal{A}(\mathbf{L}) = \mathbf{q}, \quad \|\mathbf{q} - \mathbf{b}\|_\infty \leq \delta, \quad \mathbf{P} \succeq 0. && (3.8) \end{aligned}$$

This approach can be viewed as an instance of the Douglas-Rachford variable splitting method in convex programming [102]. Next, we relax the linear constraints and form an *augmented Lagrangian* of the above problem (3.7) as follows:

$$\begin{aligned} \text{minimize} \quad & \|\mathbf{P}\|_* + \frac{\beta_1}{2}\|\mathbf{P} - \mathbf{L} - \mathbf{\Lambda}\|_F^2 + \frac{\beta_2}{2}\|\mathcal{A}(\mathbf{L}) - \mathbf{q} - \boldsymbol{\omega}\|_2^2 \\ \text{subject to} \quad & \|\mathbf{q} - \mathbf{b}\|_\infty \leq \delta. \end{aligned} \tag{3.9}$$

Here, the symmetric matrix $\mathbf{\Lambda} \in \mathbb{S}^{N\times N}$ and vector $\boldsymbol{\omega} \in \mathbb{R}^{Q'}$ represent the Lagrange multipliers. The optimization in (3.9) is carried out over the variables $\mathbf{P}, \mathbf{L} \in \mathbb{S}^{N\times N}$, and $\mathbf{q} \in \mathbb{R}^{Q'}$, and $\mathbf{\Lambda}, \boldsymbol{\omega}$ are iteratively updated as well. Instead of jointly optimizing over all three variables, we optimize the variables one at a time while keep the others fixed. Therefore, we can solve the optimization (3.9) via a sequence of three sub-problems, each of which admits an efficient solution. Let the subscript k denote the estimate of a variable at the k^{th} iteration. The following steps are performed until convergence.

1. **Update q**: Isolating the terms that involve $\mathbf{q}$, we obtain a new estimate $\mathbf{q}_{k+1}$ as the solution of the constrained optimization problem

$$\mathbf{q}_{k+1} \leftarrow \arg\min_{\mathbf{q}} \frac{\beta_2}{2}\|\mathcal{A}(\mathbf{L}_k) - \boldsymbol{\omega}_k - \mathbf{q}\|^2, \quad \text{s.t. } \|\mathbf{q} - \mathbf{b}\|_\infty \leq \delta.$$

   This problem has a closed form solution using a component-wise truncation procedure for the entries in $\mathbf{q}$. Denote $\mathbf{z} = \mathcal{A}(\mathbf{L}^k) - \boldsymbol{\omega}^k - \mathbf{b}$. Then, it is easily seen that

$$\mathbf{q}_{k+1} = \mathbf{b} + \text{sign}(\mathbf{z}) \cdot \max(|\mathbf{z}|, \delta), \tag{3.10}$$

Algorithm 3 NuMax

Inputs: Secant set $\mathcal{S}(\mathcal{X}) = \{\mathbf{v}_i\}_{i=1}^{Q'}$, isometry constant δ
Parameters: Weights β_1, β_2, step size η
Output: Symmetric PSD matrix $\widehat{\mathbf{P}}$
Initialize: $\mathbf{P}_0, \mathbf{L}_0, \boldsymbol{\omega}_0, \mathbf{q}_0, k \leftarrow 0$, $\mathbf{b} \leftarrow \mathbf{1}_{Q'}$, set $\mathcal{A} : \mathbf{X} \mapsto \{\mathbf{v}_i^T \mathbf{X} \mathbf{v}_i\}_{i=1}^{Q'}$
while not converged **do**
$k \leftarrow k + 1$
$\mathbf{z} \leftarrow \mathcal{A}(\mathbf{L}^k) - \boldsymbol{\omega}^k - \mathbf{b}$
$\mathbf{q}_{k+1} \leftarrow \mathbf{b} + \text{sign}(\mathbf{z}) \cdot \max(|\mathbf{z}|, \delta)$ {Truncation}
$\mathbf{P}' \leftarrow \mathbf{L}_k + \boldsymbol{\Lambda}_k, \quad \mathbf{P}' = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T$
$\mathbf{P}_{k+1} \leftarrow \mathbf{U}\mathcal{D}_\alpha(\boldsymbol{\Sigma})\mathbf{V}^T$ {Singular value shrinkage}
$\mathbf{Z} \leftarrow \beta_2 \mathcal{A}^*(\mathbf{q}_{k+1} + \boldsymbol{\omega}_k)$
$\mathbf{Z}' \leftarrow \beta_1(\mathbf{P}_k - \boldsymbol{\Lambda}_k)$
$\mathbf{L}_{k+1} \leftarrow \beta_2(\mathcal{A}^*\mathcal{A} + I)^\dagger(\mathbf{Z} + \mathbf{Z}')$ {Least squares}
$\boldsymbol{\Lambda}_{k+1} \leftarrow \boldsymbol{\Lambda}_k - \eta(\mathbf{P}_k - \mathbf{L}_k)$
$\boldsymbol{\omega}_{k+1} \leftarrow \boldsymbol{\omega}_k - \eta(\mathcal{A}(\mathbf{L}_k) - \mathbf{q}_k)$ {Update Lagrange multipliers}
end while
return $\widehat{\mathbf{P}} \leftarrow \mathbf{P}_k$

where the sign and max operators are applied component-wise. Therefore, this step can be performed in $\mathcal{O}(Q')$ operations.

2. **Update P**: Isolating the terms that involve $\mathbf{P}$, we obtain a new estimate $\mathbf{P}_{k+1}$ as the solution of the constrained optimization problem

$$\mathbf{P}_{k+1} \leftarrow \arg\min_{\mathbf{P}} \|\mathbf{P}\|_* + \frac{\beta_1}{2}\|\mathbf{P} - \mathbf{L}_k - \boldsymbol{\Lambda}_k\|_F^2, \quad \text{s.t. } \mathbf{P} \succeq 0.$$

This problem also admits an efficient closed form solution via the *singular value shrinkage operator* [103]. Denote $\mathbf{P}' = \mathbf{L}_k + \boldsymbol{\Lambda}_k$ and perform the SVD $\mathbf{P}' = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T$, where $\boldsymbol{\Sigma} = \text{diag}(\boldsymbol{\sigma})$. Then, the optimum $\mathbf{P}_{k+1}$ can be expressed as:

$$\mathbf{P}_{k+1} = \mathbf{U}\mathcal{D}_\alpha(\boldsymbol{\Sigma})\mathbf{V}^T, \quad \mathcal{D}_\alpha(\boldsymbol{\Sigma}) = \text{diag}(\{(\sigma_i - \alpha)_+\}), \tag{3.11}$$

where $\alpha = \frac{1}{\beta}$ and t_+ represents the positive part of t, i.e., $t_+ = \max(t, 0)$. The dominant computational cost for this update is incurred by performing the SVD of $\mathbf{P}' \in \mathbb{S}^{N \times N}$, and in general this step can be carried out in $\mathcal{O}(N^3)$ operations. This step can potentially be made even faster by using fast, approximate methods for obtaining the SVD of a matrix [104].

3. **Update L**: Isolating the terms that involve $\mathbf{L}$, we obtain a new estimate $\mathbf{L}_{k+1}$ as the solution of the unconstrained optimization problem

$$\mathbf{L}_{k+1} \leftarrow \arg\min_{\mathbf{L}} \frac{\beta_1}{2}\|\mathbf{P}_k - \mathbf{L} - \boldsymbol{\Lambda}_j\|_F^2 + \frac{\beta_2}{2}\|\mathcal{A}(\mathbf{L}) - \mathbf{q}_{k+1} - \boldsymbol{\omega}_k\|_2^2 \tag{3.12}$$

Algorithm 4 NuMax-CG

Inputs: Secant set $\mathcal{S} = \{\mathbf{v}_i\}_{i=1}^{Q'}$, isometry constant δ, the NuMax algorithm
Parameters: Size of selected secant sets Q'', Q'''
Output: Symmetric PSD matrix $\widehat{\mathbf{P}}$
Initialize: Select a subset of Q'' secants $\mathcal{S}_0$, set $\mathcal{A} : \mathbf{X} \mapsto \{\mathbf{v}_i^T \mathbf{X} \mathbf{v}_i\}_{i=1}^{Q''}$
Obtain initial estimate $\mathbf{P} \leftarrow \text{NuMax}(\mathcal{S}_0, \delta)$
while not converged **do**
$\widehat{\mathcal{S}} \leftarrow \{\mathbf{v}_i \in \mathcal{S}_0 : |\mathbf{v}_i^T \mathbf{P} \mathbf{v}_i - 1| = \delta\}$ {Retain active constraints}
$\mathcal{S}_1 \leftarrow \{\mathbf{v}_i \in \mathcal{S} : \mathbf{v}_i \notin \mathcal{S}_0\}_{i=1}^{Q'''}$ {Select additional test secants}
$\widehat{\mathcal{S}} \leftarrow \widehat{\mathcal{S}} \bigcup \{\mathbf{v}_i \in \mathcal{S}_1 : |\mathbf{v}_i^T \mathbf{P} \mathbf{v}_i - 1| \geq \delta\}$ {Secants that violate constraints}
$\mathbf{P} \leftarrow \text{NuMax}(\widehat{\mathcal{S}}, \delta)$ {Update estimate}
$\mathcal{S}_0 \leftarrow \widehat{\mathcal{S}}$
end while
return $\widehat{\mathbf{P}} \leftarrow \mathbf{P}$

This is a least-squares problem and the minimum is achieved by solving the linear system of equations:

$$\beta_1(\mathbf{P}_k - \mathbf{L} - \mathbf{\Lambda}_j) = \beta_2 \mathcal{A}^*(\mathcal{A}(\mathbf{L}) - \mathbf{q}_{k+1} - \boldsymbol{\omega}_k), \tag{3.13}$$

where $\mathcal{A}^*$ represents the adjoint of $\mathcal{A}$. The dominant cost in this step arises due to the linear operator $\mathcal{A}^*\mathcal{A}$. A single application of this operator incurs a complexity of $\mathcal{O}\left(N^2 Q'^2\right)$. The least-squares solution to (3.13) can be calculated using a number of existing methods for solving large-scale linear equations, including Conjugate Gradients [105] and L-BFGS [106].

4. **Update $\mathbf{\Lambda}, \boldsymbol{\omega}$**: Finally, as standard in Augmented Lagrange methods, we update the parameters $\mathbf{\Lambda}, \boldsymbol{\omega}$ according to the equations:

$$\mathbf{\Lambda}_{k+1} \leftarrow \mathbf{\Lambda}_k - \eta(\mathbf{P}_k - \mathbf{L}_k), \quad \boldsymbol{\omega}_{k+1} \leftarrow \boldsymbol{\omega}_k - \eta(\mathcal{A}(\mathbf{L}_k) - \mathbf{q}_k).$$

The overall NuMax method is summarized in pseudocode form in Algorithm 3. The convergence properties of NuMax, both in terms of precision as well as speed, are affected by the user-defined parameters η, β_1, and β_2. In all our experiments below in Section 3.6, we set $\eta = 1.618, \beta_1 = \beta_2 = 1$.

3.5.2 Column generation

Our proposed approach NuMax (Alg. 3) helps dramatically decrease the time-complexity of solving the optimization problem (3.4). However, for a problem with Q' input secants, the memory complexity of NuMax still remains $\mathcal{O}\left(Q'^2\right)$ and this can be prohibitive in applications involving millions (or billions) of secants. We now develop a heuristic optimization method that only approximately solves (3.4), but that scales very well to such problem sizes.

The key idea behind our proposed method is based on the Karush-Kuhn-Tucker (KKT) conditions describing the optimum of (3.4). Recall that (3.4) consists of optimizing a linear objective

subject to inequality constraints over the cone of PSD matrices. Suppose that strong duality holds, i.e., the primal and dual optimal values of (3.4) are equal. Then, classical results in optimization theory [107] states that *complementary slackness* holds and that the optimal solution is entirely specified by the set of those constraints that hold with equality; such constraints are also known as *active* constraints.

We therefore propose a simple, greedy method to rapidly find the active constraints for a problem of the form (3.4). We prescribe the following steps:

1. Solve (3.4) with only a small subset $\mathcal{S}_0$ of the input secants $\mathcal{S}(\mathcal{X})$ using NuMax (Alg. 3) to obtain an initial estimate $\mathbf{P}$. Identify the set $\widehat{\mathcal{S}}$ of secants that correspond to active constraints, i.e,
$$\widehat{\mathcal{S}} \leftarrow \{\mathbf{v}_i \in \mathcal{S}_0 : |\mathbf{v}_i^T \mathbf{P} \mathbf{v}_i - 1| = \delta\}.$$
2. Select a small number of additional secants $\mathcal{S}_1 \subset \mathcal{S}$ that were not selected previously and identify all the secants among S_1 that *violate* the constraints at the current estimate $\mathbf{P}$. Append these secants to the set of active constraints $\widehat{\mathcal{S}}$ to obtain an augmented set $\widehat{\mathcal{S}}$.
$$\widehat{\mathcal{S}} \leftarrow \widehat{\mathcal{S}} \bigcup \{\mathbf{v}_i \in \mathcal{S}_1 : |\mathbf{v}_i^T \mathbf{P} \mathbf{v}_i - 1| \geq \delta\}.$$
3. Solve (3.4) with the augmented set $\widehat{\mathcal{S}}$ using NuMax to obtain an new estimate $\mathbf{P}$.
4. Identify the secants that correspond to active constraints. Repeat Steps 2 and 3, and so on, until convergence in the estimated optimal matrix $\mathbf{P}$.

Therefore, instead of performing a large numerical optimization procedure on the entire set of secants $\mathcal{S}(\mathcal{X})$, we only perform a sequence of optimization procedures on small subsets of $\mathcal{S}(\mathcal{X})$. This approach is analogous to the *column generation* (CG) method used to solve very large-scale linear programs. Therefore, we dub our overall algorithm *NuMax-CG*; this algorithm is listed in pseudocode form in Algorithm 4

Another key benefit of our proposed NuMax-CG method, the set of secants upon which NuMax acts upon within each iteration never needs to be explicitly stored in memory and can in fact be generated *on the fly*. This can potentially lead to significant improvements in terms of memory complexity of the overall procedure. An important caveat is that we are no longer guaranteed to converge to the optimal solution of (3.4); nevertheless, as we see below in Section 3.6, our proposed NuMax-CG algorithm yields excellent results on massively sized real-world datasets.

3.6 Experiments

3.6.1 Linear low-dimensional embeddings

We illustrate the performance of our proposed NuMax algorithm for generating low-dimensional near-isometric embeddings. First, we consider a synthetic manifold $\mathcal{M}$ that comprises 16×16 images of shifts of a white square on a black background (Fig. 3.1) so that $N = 256$. The degrees of freedom for each image are simply the 2D coordinates of the center of the square and hence

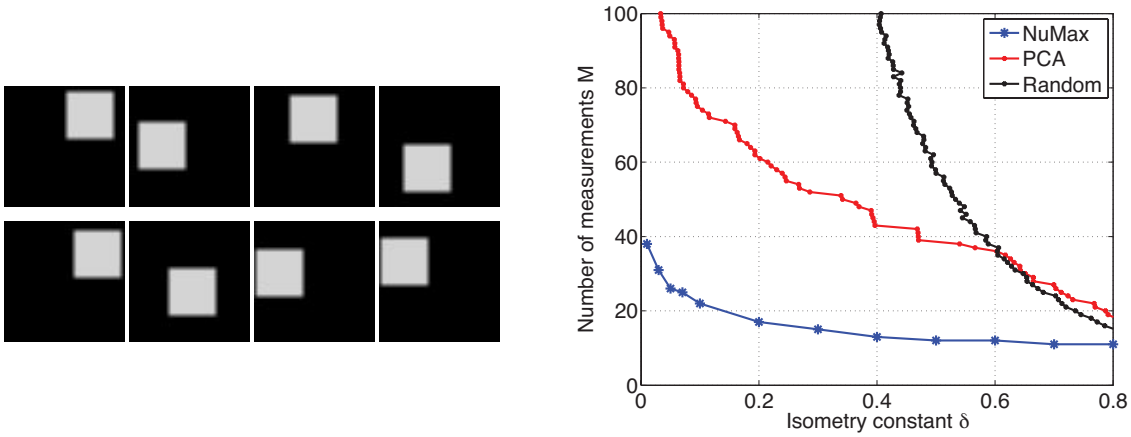


Figure 3.1: *(left) Example images from the manifold of translating squares. (right) Empirical isometry constant δ vs. number of measurements M using various types of embeddings. The SDP secant embedding approach ensures global approximate isometry using the fewest number of measurements.*

the dimension of the manifold is 2. We construct a training set $\mathcal{S}(\mathcal{X})$ of $Q' = 1000$ secants by randomly sampling pairs of images from this manifold and normalize the secants using 3.2, Then, we solve (3.4) with desired isometry constant δ using Algorithm 3 to obtain a positive semi-definite symmetric matrix $\mathbf{P}^*$. We measure the rank of $\mathbf{P}^*$ and denote it by M.

Next, we perform an empirical estimation of isometry constants obtained via PCA. We achieve this by projecting the secants on to M PCA basis functions learned on the secant set $\mathcal{S}(\mathcal{X})$ and calculating the norms of the projected secants. The worst-case deviation from unity gives the estimate of the isometry constant. We perform a similar isometry constant calculation using random Gaussian projections.

Figure 3.1 plots the variation of the number of measurements M with the isometry constant δ. Figure 3.1 can be viewed as analogous to the *rate-distortion curve* commonly studied in Shannon information theory; here, δ represents the distortion and the undersampling factor M/N represents the compression rate. We observe that our proposed NuMax embedding $\mathbf{\Psi}$ achieves the desired isometry on the secants using the fewest number of measurements. Interestingly, both the NuMax embeddings as well as the PCA embeddings far outperform the random projection approach.

Figure 3.2 demonstrates the computational efficiency of the NuMax algorithm. We generate N-pixel images of translating disks, contruct Q' training secants, and run NuMax until convergence. For a problem size with $Q' = 4000$ and $N = 1024$ (i.e., over 1 million variables), we observe that NuMax takes only a few minutes to converge. On the other hand, traditional interior-point methods (such as those employed in CVX [85]) takes over 24 hours to produce the same answer. We also observe a *linear* relationship in the runtime of NuMax with the number of input secants Q'. This represents a significant improvement over the runtimes of best-known interior-point methods, whose complexity is known to be $\mathcal{O}\left(Q'^2\right)$.

We repeat this experiment on a real-world dataset. The MNIST dataset consists of a large

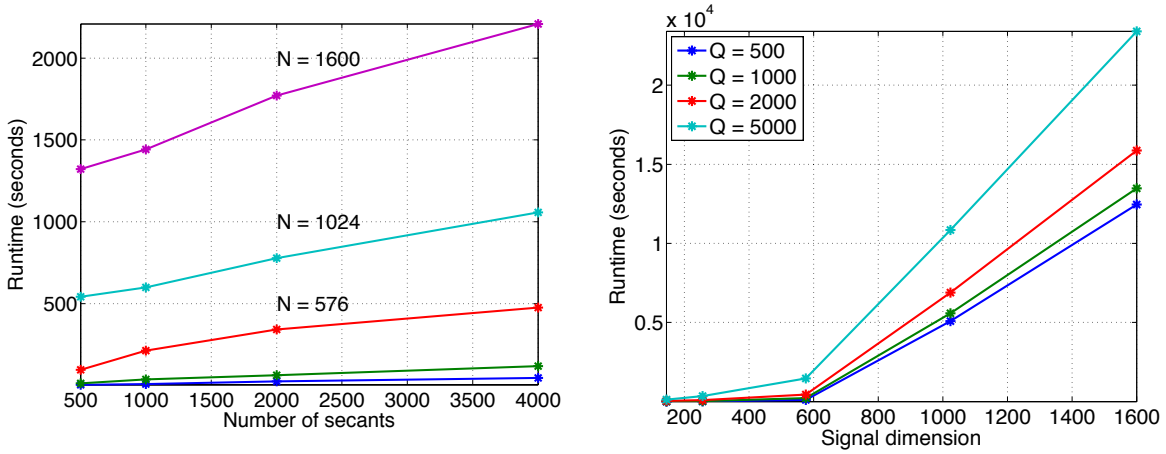


Figure 3.2: *Empirical characterization of computational complexity of Algorithm 3. (left) Variation of average runtime (in seconds) with number of secants Q. (right) Variation of average runtime (in seconds) with signal dimension N.*

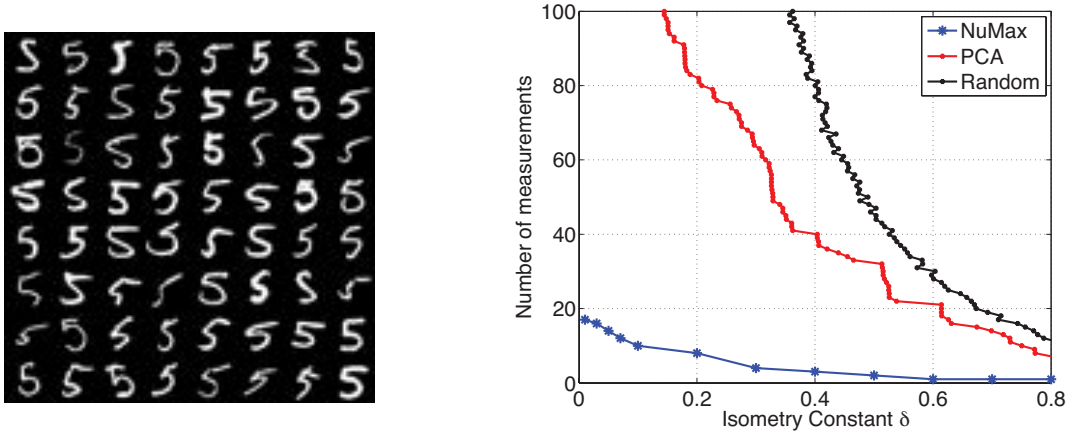


Figure 3.3: *(left) Example images from the MNIST dataset.(right) Empirical isometry constant δ vs. number of measurements M using various types of embeddings. The NuMax secant embedding approach ensures global approximate isometry using the fewest number of measurements.*

database of digital images of handwritten digits and is commonly used as a benchmark for various machine learning algorithms. However, the images themselves exhibit dramatic variations (see Fig. 3.3) and presumably the set of all images forms a highly nonlinear, non-differentiable manifold. We construct a training dataset $\mathcal{S}(\mathcal{X})$ comprising $Q' = 3000$ secants and estimate the variation of the isometry constant δ with the number of measurements M. The results of this experiment are plotted in Fig. 3.3. Once again, we observe that our proposed approach provides the best linear embedding for a given value of δ in terms of reduced dimensionality.

3.6.2 Approximate Nearest Neighbors (ANN)

The notion of *nearest neighbors* (NN) is vital to numerous problems in estimation, classification, and regression [108]. Suppose that a large dataset of training examples is available. Then, given a new (query) data point, NN-based machine learning techniques typically identify the k points in the training dataset closest to the query point and use these identified points for further processing. Due to its conceptual simplicity, the use of nearest neighbors has become ubiquitous in machine learning.

Suppose that the data points are modeled as elements of a vector space. As the dimension N of the data grows, the computational cost of finding the k nearest neighbors becomes challenging [109]. To counter this challenge, as opposed to computing nearest neighbors of the query data point, one can instead construct a near-isometric embedding of the data into an M-dimensional space and only attempt to estimate *approximate nearest neighbors* (ANN) in the embedded space. By carefully controlling the distortion in distances caused by the lower-dimensional embedding, efficient inference techniques can be performed with little loss in performance. The ANN principle forms the core of *locality sensitive hashing* (LSH), a popular technique for high-dimensional pattern recognition and information retrieval [110, 111].

Given a fixed dataset, the time complexity of a particular ANN method directly depends upon the dimension M of the embedded space; the smaller the embedding dimension, the faster the ANN method. Most existing ANN methods (including LSH) either involve computing a randomized linear dimensionality reduction of the data, or involve a PCA decomposition of the data. In contrast, we immediately observe that our proposed NuMax algorithm provides a linear near-isometric embedding that achieves a given distortion δ while *minimizing* M. In other words, our algorithm can potentially enable far more efficient ANN computations than traditional data sets.

We test the efficiency of our approach on a set of $Q = 4000$ images taken from the LabelMe database [112]. This database consists of high resolution photographs of both indoor and outdoor scenes (see Fig. 3.4) for some examples. We compute GIST feature descriptors [113] for every image. In our case, the GIST descriptors are vectors of size $N = 512$ that coarsely express the dominant spatial statistics of the scene; such descriptors have been shown to be very useful for image retrieval purposes. Therefore our "ground truth" data consists of a matrix of size $N \times Q$.

Since the number of pairwise secants in this case is extremely high ($Q^2 = 16 \times 10^6$), we use NuMax-CG (Alg. 4) to estimate the linear embedding for a given distortion parameter δ. We fix a desired distortion parameter δ and use Alg. 4 to estimate the matrix $\mathbf{P} = \mathbf{\Psi}^T\mathbf{\Psi}$, where $\mathbf{\Psi}$ represents the linear embedding. We record M, the rank of the optimal $\mathbf{P}$ and for comparison purposes, also compute M-dimensional random linear projections of the data, as well as the best M-term PCA decomposition of the data. Subsequent ANN computations for a set of 1000 test query points are performed in the corresponding M-dimensional space.

Figure 3.5 displays the benefits in ANN computations using the linear embedding generated by our proposed NuMax-CG algorithm. For a given neighborhood size k, we plot the fraction of k-nearest neighbors computed using the full (ground truth) N-dimensional data that are also k-nearest neighbors in the corresponding M-dimensional embedding. We observe from Fig. 3.5 that the linear embedding obtained by the NuMax-CG algorithm provides the best embedding results for a wide range of measurements M and neighborhood sizes k. In particular, for embedding

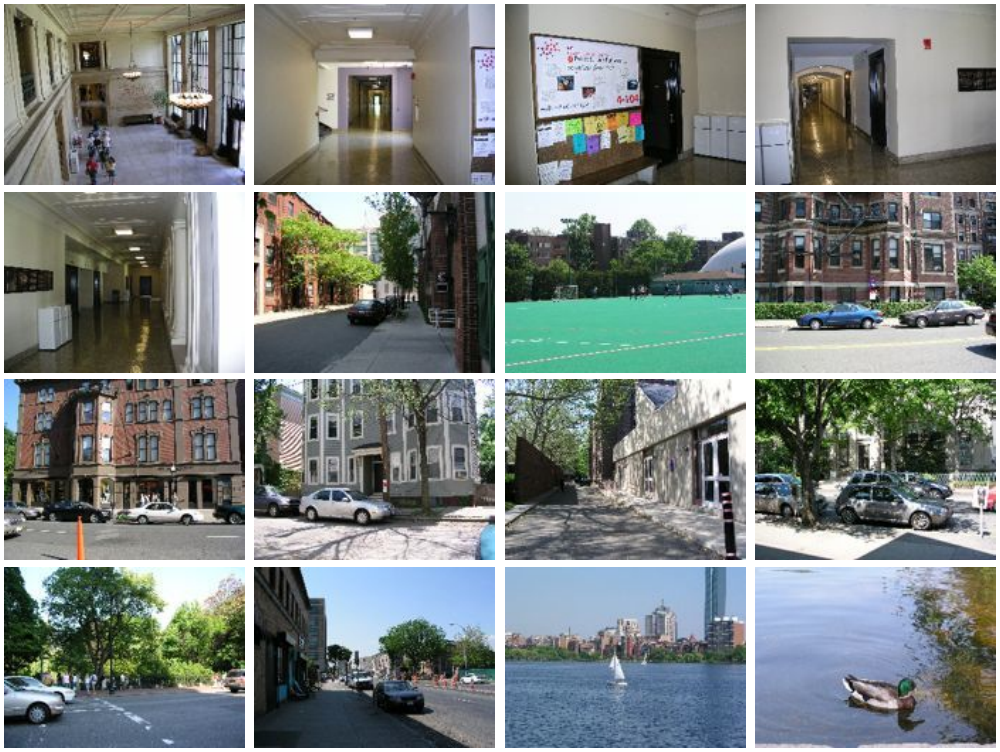


Figure 3.4: *Example images from the LabelMe dataset.*

dimensions of $M > 45$, we observe that our proposed embedding outperforms both PCA as well as random projections for all values of k by a significant margin; in fact, our approach retains nearly twice as many nearest neighbors

3.6.3 Supervised binary classification

We consider a supervised classification problem, where our classes consist of binary images of shifts of a translating disk and a translating square (some examples are shown in Fig. 3.6). We construct a training dataset of $Q' = 2000$ *inter-class* secants and obtain a measurement matrix $\mathbf{\Psi}_{\text{inter}}$ via the NuMax algorithm. Using a small number of measurements of a test signal, we estimate the class label using a Generalized Maximum Likelihood Classification (GMLC) approach following the framework in [114]. We repeat the above classification experiment using measurement basis functions learned via PCA on the inter-class secants, as well as random projections. Figure 3.6(c) plots the variation of the number of measurements M vs. the probability of error. Again, we observe that our proposed SDP formulation outperforms PCA as well as random projections.

Finally, we test our approach on a more challenging dataset. We consider the CVDomes dataset, a collection of simulated X-band radar scattering databases for civilian vehicles [115]. The database for each vehicle class consists of (complex-valued) short-time signatures of length $N = 128$ for the entire 360° azimuth, sampled every 0.0625°. We model each of the two classes as a one-dimensional

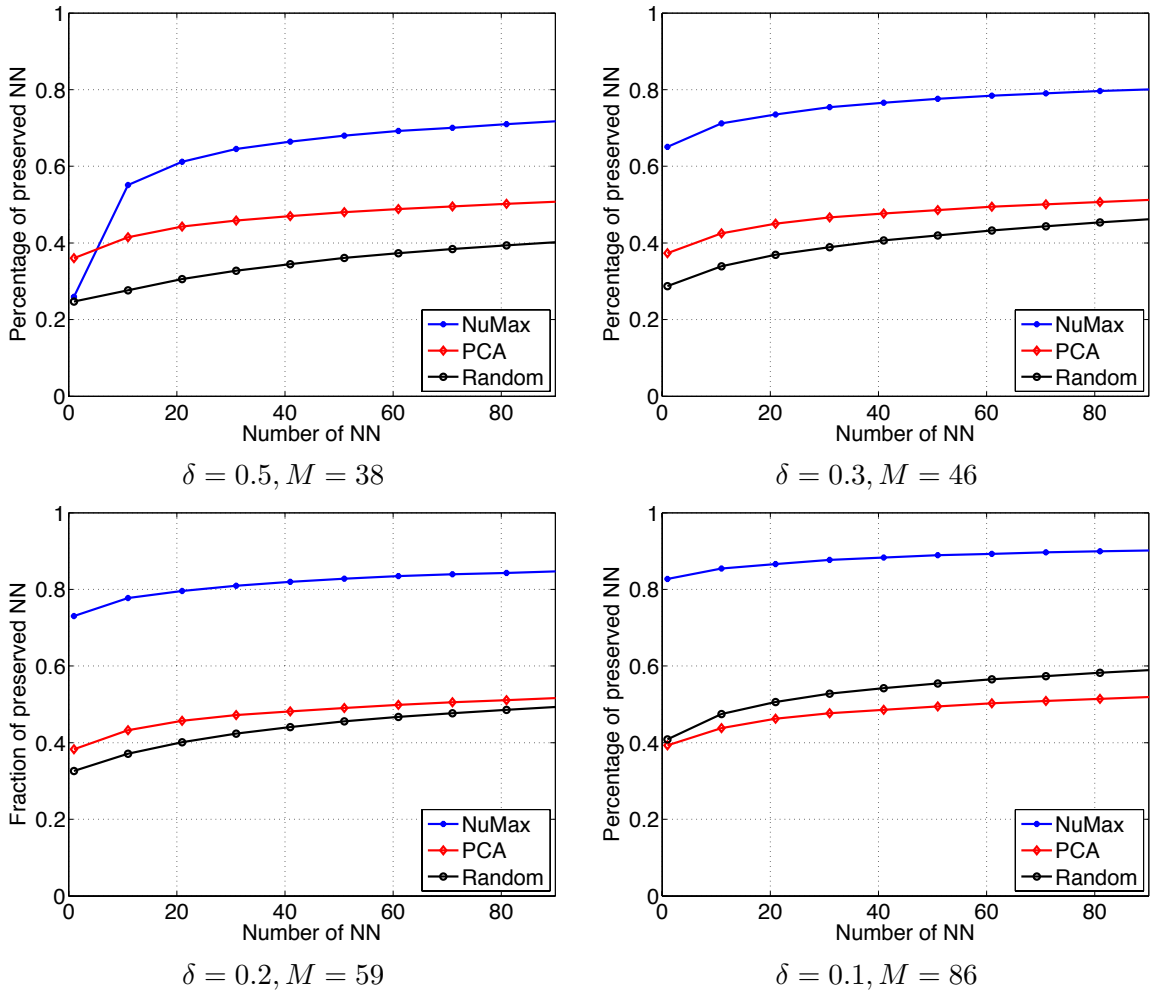


Figure 3.5: *Approximate Nearest Neighbors (ANN) for the LabelMe dataset using various linear embedding methods. A set of 4000 images are chosen and GIST features of size $N = 512$ are computed for every image; these represent the data vectors. For a given number of nearest neighbors k, we plot the average fraction of k-nearest neighbors that are retained in an M-dimensional embedding relative to the full N-dimensional data. Our proposed NuMax algorithm provides the best embedding results for a wide range of measurements M and neighborhood sizes k.*

manifold, construct $Q' = 2000$ training inter-class secants and compute the measurement matrix $\mathbf{\Psi}_{\text{inter}}$ using our proposed NuMax algorithm. Additionally, we obtain a different measurement matrix $\mathbf{\Psi}_{\text{joint}}$ from a training dataset of $Q' = 2000$ vectors that comprise both inter- and intra-class secants. Given a new test signal, we acquire M linear measurements and perform maximum likelihood classification. From Fig. 3.7, we infer that the SDP approach of optimizing (3.4) over the inter-class secants yields the best classification performance among the differerent measurement techniques. In particular, $\mathbf{\Psi}_{\text{inter}}$ produces the best classification performance, proving the potential

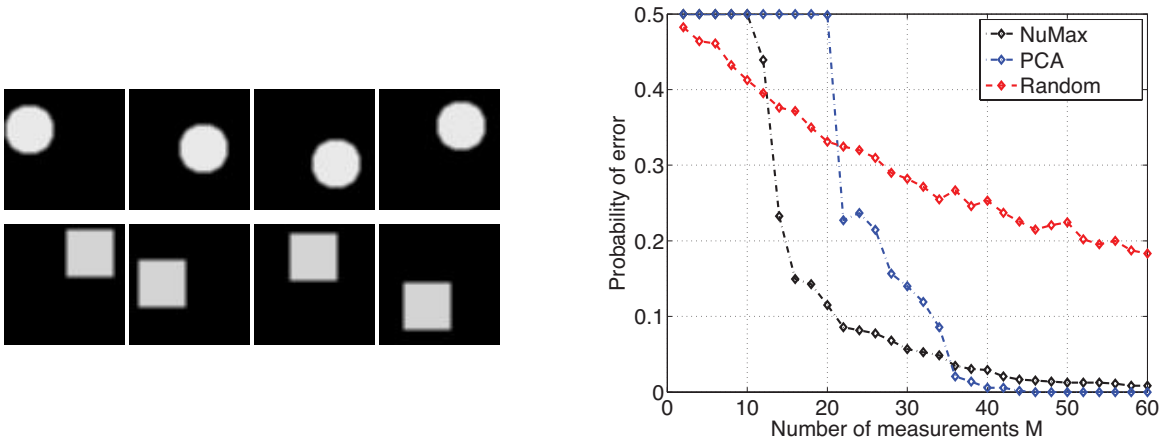


Figure 3.6: *Binary classification from low-dimensional linear embeddings. (left) The signals of interest comprise shifted images of a white disk/square on a black background. We observe M linear measurements of a test image using different matrices, and classify the observed samples using a GMLC approach. (right) Observed probability of classification error as a function of M. Our NuMax Secant approach yields high classification rates using very few measurements.*

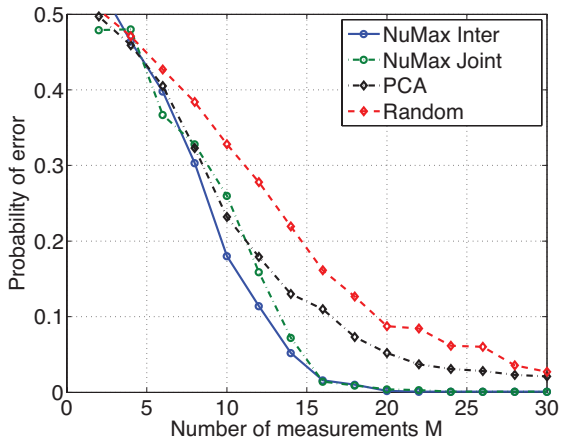


Figure 3.7: *Classification of a Toyota Camry versus a Nissan Maxima using M linear measurements of length-128 radar signatures. The SDP approach produces $> 95\%$ classification rates using only a small number of measurements $M \leq 15$.*

benefits of our approach.

3.7 Discussion

We have taken some initial steps towards a new method for constructing a near-isometric, low-dimensional embedding of manifold-modeled signals and images. Our proposed linear embedding

preserves the norms of all pairwise secants of a high-resolution sampling of the manifold and is calculated via a novel semi-definite programming (SDP) formulation (3.4). We propose an efficient ADMM-type algorithm, that we dub NuMax, for efficiently constructing the embedding from training data. Our method can be easily adapted to perform more complicated inference tasks, such as binary supervised classification.

Our proposed approach can be extended in a number of scenarios. More generally, the NuMax algorithm is directly applicable to affine rank minimization problems (of the type described in Section 2.4) that are governed by max-norm constraints. Such formulations are naturally encountered in fields such as system identification and optics. We discuss this further in Chapter 7.

Some key challenges remain. First, our proposed approach relies on the efficiency of the nuclear norm as a valid proxy for the matrix rank in the objective function in (3.4). A natural question is to examine under what conditions the optimum of the convex relaxation (3.4) equals the optimum of the nonconvex problem (3.3). Second, while the convergence rates of NuMax is shown to empirically shown to be far better than traditional methods, a theoretical validation of this statement can be a challenging task. In Chapter 7, we revisit some of these questions.

CHAPTER 4

Signal Recovery on Incoherent Manifolds

4.1 Setup

In this chapter, we consider the problem of *reconstruction* of manifold-modeled signals from noisy linear samples. Estimation of an unknown signal from linear observations is a core problem in signal processing, statistics, and information theory. Particular energy has been invested in problem instances where the available information is *limited and noisy* and where the signals of interest possess a *low-dimensional* geometric structure. Indeed, focused efforts on certain instances of the linear inverse problem framework have spawned entire research subfields, encompassing both theoretical and algorithmic advances. Examples include signal separation and morphological component analysis [44, 118]; sparse approximation and compressive sensing [20, 21, 119]; affine rank minimization [58]; and robust principal component analysis [59, 120].

We first consider a general instance of our proposed framework. Suppose that the signal of interest $\mathbf{x}^*$ can be written as the sum of two constituent signals $\mathbf{a}^* \in \mathcal{A}$ and $\mathbf{b}^* \in \mathcal{B}$, where $\mathcal{A}, \mathcal{B}$ are nonlinear, possibly non-differentiable, sub-manifolds of the signal space $\mathbb{R}^N$. Suppose that we are given access to noisy linear measurements of $\mathbf{x}^*$:

$$\mathbf{z} = \mathbf{\Phi}(\mathbf{a}^* + \mathbf{b}^*) + \mathbf{e}, \tag{4.1}$$

where $\mathbf{\Phi} \in \mathbb{R}^{M \times N}$ is the measurement matrix. Our objective is to recover the pair of signals $(\mathbf{a}^*, \mathbf{b}^*)$, and thus also $\mathbf{x}^*$, from $\mathbf{z}$. At the outset, numerous obstacles arise while trying to solve (4.1), some of which appear to be insurmountable:

1. (Identifiability I) Consider even the simplest case, where the measurements are noiseless and

This work is in collaboration with Richard G. Baraniuk [116, 117].

the measurement operator is the identity, i.e., we observe $\mathbf{x} \in \mathbb{R}^N$ such that

$$\mathbf{x} = \mathbf{a}^* + \mathbf{b}^*, \tag{4.2}$$

where $\mathbf{a}^* \in \mathcal{A}, \mathbf{b}^* \in \mathcal{B}$. This expression for $\mathbf{x}$ contains $2N$ unknowns but only N observations and hence is fundamentally ill-posed. Unless we make additional assumptions on the geometric structure of the component manifolds $\mathcal{A}$ and $\mathcal{B}$, a unique decomposition of $\mathbf{x}$ into its constituent signals $(\mathbf{a}^*, \mathbf{b}^*)$ may not exist.

2. (Identifiability II) To complicate matters, in more general situations the linear operator $\mathbf{\Phi}$ in (4.1) might have fewer rows that columns, so that $M < N$. Thus, $\mathbf{\Phi}$ possesses a nontrivial nullspace. Indeed, we are particularly interested in cases where $M \ll N$, in which case the nullspace of Φ is extremely large relative to the ambient space. This further obscures the issue of identifiability of the ordered pair $(\mathbf{a}^*, \mathbf{b}^*)$, given the available observations $\mathbf{z}$.

3. (Nonconvexity) Even if the above two identifiability issues were resolved, the manifolds $\mathcal{A}, \mathcal{B}$ might be extremely nonconvex, or even non-differentiable. Thus, classical numerical methods, such as Newton's method or steepest descent, cannot be successfully applied; neither can the litany of convex optimization methods that have been specially designed for linear inverse problems with certain types of signal priors [44, 58].

In this paper, we propose a simple method to recover the component signals $(\mathbf{a}^*, \mathbf{b}^*)$ from $\mathbf{z}$ in (4.1). We dub our method *Successive Projections onto INcoherent manifolds* (SPIN) (see Algorithm 5) Despite the highly nonconvex nature of the problem and the possibility of underdetermined measurements, SPIN *provably* recovers the signal components $(\mathbf{a}^*, \mathbf{b}^*)$. For this to hold true, we will require that (i) the signal manifolds $\mathcal{A}, \mathcal{B}$ are *incoherent* in the sense that the secants of $\mathcal{A}$ are almost orthogonal to the secants of $\mathcal{B}$; and (ii) the measurement operator $\mathbf{\Phi}$ satisfies a certain *restricted isometry property* (RIP) on the secants of the direct sum manifold $\mathcal{C} = \mathcal{A} \oplus \mathcal{B}$. We will formally define these conditions in Section 4.3. We prove the following theoretical statement below in Section 4.4.

Theorem 1 (Signal recovery)**.** *Let $\mathcal{A}, \mathcal{B}$ be incoherent manifolds in $\mathbb{R}^N$. Let $\mathbf{\Phi}$ be a measurement matrix that satisfies the RIP on the direct sum $\mathcal{C} = \mathcal{A} \oplus \mathcal{B}$. Suppose we observe linear measurements $\mathbf{z} = \mathbf{\Phi}(\mathbf{a}^* + \mathbf{b}^*)$, where $\mathbf{a}^* \in \mathcal{A}$ and $\mathbf{b}^* \in \mathcal{B}$. Then, given any precision parameter $\nu > 0$, there exists a positive integer T_ν and an iterative algorithm that outputs a sequence of iterates $(\mathbf{a}_k, \mathbf{b}_k) \in \mathcal{A} \times B, k = 1, 2, \ldots$ such that $\max\{\|\mathbf{a}_k - \mathbf{a}^*\|_2, \|\mathbf{b}_k - \mathbf{b}^*\|_2\} \leq 1.5\nu$ for all $k > T_\nu$.*

Our proposed algorithm (SPIN) is iterative in nature. Each iteration consists of three steps: computation of the gradient of the error function $\psi(\mathbf{a}, \mathbf{b}) = \frac{1}{2}\|\mathbf{z} - \mathbf{\Phi}(\mathbf{a} + \mathbf{b})\|_2^2$, forming signal proxies for $\mathbf{a}$ and $\mathbf{b}$, and orthogonally projecting the proxies onto the manifolds $\mathcal{A}$ and $\mathcal{B}$. The projection operators onto the component manifolds play a crucial role in algorithm stability and performance; some manifolds admit stable, efficient projection operators while others do not. We discuss this in detail in Section 4.4. Additionally, we demonstrate that SPIN is stable to measurement noise (the quantity $\mathbf{e}$ in (4.1)) as well as numerical inaccuracies (such as finite precision arithmetic).

4.2 Related Work

The core essence of our proposed approach has been extensively studied in a number of different contexts. Methods such as Projected Landweber iterations [121], iterative hard thresholding (IHT) [48], and singular value projection (SVP) [96] are all instances of the same basic framework. SPIN subsumes and generalizes these methods. In particular, SPIN is an iterative projected gradient method with the same basic approach as two recent signal recovery algorithms — Gradient Descent with Sparsification (GraDeS) [122], and Manifold Iterative Pursuit (MIP) [123]. We generalize these approaches to situations where the signal of interest is a linear mixture of signals arising from a pair of nonlinear manifolds. Due to the particular structure of our setting, SPIN consists of *two* projection steps (instead of one), and the analysis is more involved (see Section A.1). We also explore the interplay between the geometric structure of the component manifolds, the linear measurement operator, and the stability of the recovery algorithm.

SPIN exhibits a strong geometric convergence rate comparable to many state-of-the-art first-order methods [48, 96], despite the nonlinear and nonconvex nature of the reconstruction problem. We duly note that, for the case of certain special manifolds, sophisticated higher-order recovery methods with stronger stability guarantees have been proposed (e.g., approximate message passing (AMP) [124] for sparse signal recovery and augmented Lagrangian multiplier (ALM) methods for low-rank matrix recovery [56]); see also [125]. However, an appealing feature of SPIN is its conceptual simplicity plus its ability to generalize to mixtures of arbitrary nonlinear manifolds, provided these manifolds satisfy certain geometric properties, as detailed in Section 4.3.

4.3 Geometric Assumptions

The analysis and proof of accuracy of SPIN (Algorithm 5) involves three core ingredients: (i) a geometric notion of *manifold incoherence* that crystallizes the approximate orthogonality between secants of submanifolds of $\mathbb{R}^N$; (ii) a *restricted isometry* condition satisfied by the measurement operator $\mathbf{\Phi}$ over the secants of a submanifold; and (iii) the availability of *projection operators* that compute the orthogonal projection of any point $\mathbf{x} \in \mathbb{R}^N$ onto a submanifold of $\mathbb{R}^N$.

4.3.1 Manifold incoherence

In linear inverse problems such as sparse signal approximation and compressive sensing, the assumption of incoherence between linear subspaces, bases, or dictionary elements is common. We introduce a nonlinear generalization of this concept. For ease of notation, let $\|\cdot\|$ denote the ℓ_2-norm. Recall from Chapter 3 that for any manifold $\mathcal{M}$, the secant manifold $\mathcal{S}(\mathcal{A})$ is the family of unit vectors $\mathbf{u}$: $\mathbf{u} = \frac{\mathbf{x}-\mathbf{x}'}{\|\mathbf{x}-\mathbf{x}'\|},\ \mathbf{x}, \mathbf{x}' \in \mathcal{M},\ \mathbf{x} \neq \mathbf{x}'$, generated by all pairs $\mathbf{x}, \mathbf{x}'$ in $\mathcal{M}$.

Definition 5. *Suppose $\mathcal{A}, \mathcal{B}$ are submanifolds of $\mathbb{R}^N$. Let*

$$\sup_{\mathbf{u} \in \mathcal{S}(\mathcal{A}),\ \mathbf{u}' \in \mathcal{S}(\mathcal{B})} \left| \langle \mathbf{u}, \mathbf{u}' \rangle \right| = \epsilon, \tag{4.3}$$

where $\mathcal{S}(\mathcal{A}), \mathcal{S}(\mathcal{B})$ are the secant manifolds of $\mathcal{A}, \mathcal{B}$ respectively. Then, $\mathcal{A}$ and $\mathcal{B}$ are called ϵ-incoherent *manifolds.*

Informally,the incoherence parameter ϵ controls the extent of "perpendicularity" between the manifolds $\mathcal{A}$ and $\mathcal{B}$. We define ϵ in terms of a supremum over sets $\mathcal{S}(\mathcal{A}), \mathcal{S}(\mathcal{B})$. Therefore, a small value of ϵ implies that *each* (normalized) secant of $\mathcal{A}$ is approximately orthogonal to *all* secants of $\mathcal{B}$. By definition, the quantity ϵ is always non-negative; further, $\epsilon \leq 1$, due to the Cauchy-Schwartz inequality.

We prove that any signal $\mathbf{x}$ belonging to the direct sum $\mathcal{A} \oplus \mathcal{B}$ can be *uniquely* decomposed into its constituent signals when the upper bound on ϵ holds with strict inequality.

Lemma 4 (Uniqueness)**.** *Suppose that $\mathcal{A}, \mathcal{B}$ are ϵ-incoherent with $0 < \epsilon < 1$. Consider $\mathbf{x} = \mathbf{a} + \mathbf{b} = \mathbf{a}' + \mathbf{b}'$, where $\mathbf{a}, \mathbf{a}' \in \mathcal{A}$ and $\mathbf{b}, \mathbf{b}' \in \mathcal{B}$. Then, $\mathbf{a} = \mathbf{a}', \mathbf{b} = \mathbf{b}'$.*

Proof. It is clear that $\|\mathbf{a} + \mathbf{b} - (\mathbf{a}' + \mathbf{b}')\|^2 = 0$, i.e.,

$$\|\mathbf{a} - \mathbf{a}'\|^2 + \|\mathbf{b} - \mathbf{b}'\|^2 = -2\langle \mathbf{a} - \mathbf{a}', \mathbf{b} - \mathbf{b}' \rangle \leq 2\left|\langle \mathbf{a} - \mathbf{a}', \mathbf{b} - \mathbf{b}' \rangle\right|.$$

However, due to the manifold incoherence assumption, the (unnormalized) secants $\mathbf{a} - \mathbf{a}'$, $\mathbf{b} - \mathbf{b}'$ obey the relation:

$$\left|\langle \mathbf{a} - \mathbf{a}', \mathbf{b} - \mathbf{b}' \rangle\right| \leq \epsilon \|\mathbf{a} - \mathbf{a}'\| \|\mathbf{b} - \mathbf{b}'\| \leq \frac{1}{2}\epsilon(\|\mathbf{a} - \mathbf{a}'\|^2 + \|\mathbf{b} - \mathbf{b}'\|^2), \tag{4.4}$$

where the last inequality follows from the relation between arithmetic and geometric means (henceforth referred to as the *AM-GM inequality*). Therefore, we have that

$$\|\mathbf{a} - \mathbf{a}'\|^2 + \|\mathbf{b} - \mathbf{b}'\|^2 \leq \epsilon \left(\|\mathbf{a} - \mathbf{a}'\|^2 + \|\mathbf{b} - \mathbf{b}'\|^2\right),$$

for $\epsilon < 1$, which is impossible unless $\mathbf{a} = \mathbf{a}', \mathbf{b} = \mathbf{b}'$. □

We can also prove the following relation between secants and direct sums of signals lying on incoherent manifolds.

Lemma 5. *Suppose that $\mathcal{A}, \mathcal{B}$ are ϵ-incoherent with $0 < \epsilon < 1$. Consider $\mathbf{x}_1 = \mathbf{a}_1 + \mathbf{b}_1, \mathbf{x}_2 = \mathbf{a}_2 + \mathbf{b}_2$, where $\mathbf{a}_1, \mathbf{a}_2 \in \mathcal{A}$ and $\mathbf{b}_1, \mathbf{b}_2 \in \mathcal{B}$. Then*

$$\left|\langle \mathbf{a}_1 - \mathbf{a}_2, \mathbf{b}_1 - \mathbf{b}_2 \rangle\right| \leq \frac{\epsilon}{2(1-\epsilon)} \|\mathbf{x}_1 - \mathbf{x}_2\|^2.$$

Proof. From (4.4), we have

$$\begin{aligned} \left|\langle \mathbf{a}_1 - \mathbf{a}_2, \mathbf{b}_1 - \mathbf{b}_2 \rangle\right| &\leq \frac{\epsilon}{2}(\|\mathbf{a}_1 - \mathbf{a}_2\|^2 + \|\mathbf{b}_1 - \mathbf{b}_2\|^2) \\ &= \frac{\epsilon}{2}\|\mathbf{a}_1 + \mathbf{b}_1 - \mathbf{a}_2 - \mathbf{b}_2\|^2 - \epsilon\langle \mathbf{a}_1 - \mathbf{a}_2, \mathbf{b}_1 - \mathbf{b}_2 \rangle \\ &\leq \frac{\epsilon}{2}\|\mathbf{x}_1 - \mathbf{x}_2\|^2 + \epsilon\left|\langle \mathbf{a}_1 - \mathbf{a}_2, \mathbf{b}_1 - \mathbf{b}_2 \rangle\right|. \end{aligned}$$

Rearranging terms, we obtain the desired result. □

4.3.2 Restricted isometry

Next, we address the situation where the measurement operator $\mathbf{\Phi} \in \mathbb{R}^{M \times N}$ contains a nontrivial nullspace, i.e., when $M < N$. We will require that $\mathbf{\Phi}$ satisfies the restricted isometry property (RIP) on the *secants* of the *direct sum* $\mathcal{C} = \mathcal{A} \oplus \mathcal{B}$. Concretely, we have that for every normalized secant $\mathbf{u}$ belonging to the secant manifold $\mathcal{S}(\mathcal{C})$, we have that

$$1 - \delta \leq \|\Phi \mathbf{u}\|^2 \leq 1 + \delta, \tag{4.5}$$

for some $\delta \in [0, 1)$. We observe that (4.5) is similar to (2.8), except that the notion of restricted isometry is defined over all unit vectors belonging to the secant manifold $\mathcal{S}(\mathcal{C})$.

As discussed in Section 2.3, the notion of restricted isometry (and its generalizations) is an important component in the analysis of many algorithms in sparse approximation, compressive sensing, and low-rank matrix recovery [20, 58]. The central result in [82] states that, under certain upper bounds on the curvature of the manifold $\mathcal{C}$, there exist probabilistic constructions of matrices $\mathbf{\Phi}$ that satisfy the RIP on $\mathcal{C}$ such that the number of rows of $\mathbf{\Phi}$ is proportional to the intrinsic dimension of $\mathcal{C}$, rather than the ambient dimension N of the signal space. We revisit this result in greater detail in Chapter 5.

4.3.3 Projections onto manifolds

Given an arbitrary nonlinear manifold $\mathcal{A} \in \mathbb{R}^N$, we define the operator $\mathcal{P}_{\mathcal{A}}(\cdot) : \mathbb{R}^N \mapsto \mathcal{A}$ as the Euclidean projection operator onto $\mathcal{A}$:

$$\mathcal{P}_{\mathcal{A}}(\mathbf{x}) = \arg\min_{\mathbf{x}' \in \mathcal{A}} \left\|\mathbf{x}' - \mathbf{x}\right\|^2. \tag{4.6}$$

Informally, given an arbitrary vector $\mathbf{x} \in \mathbb{R}^N$, the operator $\mathcal{P}_{\mathcal{A}}(x)$ returns the point on the manifold $\mathcal{A}$ that is "closest" to $\mathbf{x}$, where closeness is measured in terms of the Euclidean norm. Observe that for arbitrary nonconvex manifolds $\mathcal{A}$, the above minimization problem (4.6) may not yield a unique optimum. Technically, therefore, $\mathcal{P}_{\mathcal{A}}(\mathbf{x})$ is a set-valued operator. For ease of exposition, $\mathcal{P}_{\mathcal{A}}(\mathbf{x})$ will henceforth refer to *any* arbitrarily chosen element of the set of signals that minimize the ℓ_2-error in (4.6).

The projection operator $\mathcal{P}_{\mathcal{A}}(\cdot)$ plays a crucial role in the development of our proposed signal recovery algorithm in Section 4.4. Note that in a number of applications, $\mathcal{P}_{\mathcal{A}}(\cdot)$ may be quite difficult to compute exactly. The reasons for this might be intrinsic to the application (such as the nonconvex, non-differentiable structure of $\mathcal{A}$), or might be due to extrinsic constraints (such as finite-precision arithmetic). Therefore, following the lead of [123], we also define a γ-approximate projection operator onto $\mathcal{A}$:

$$\mathbf{x}' = \mathcal{P}^{\gamma}_{\mathcal{A}}(\mathbf{x}) \implies \mathbf{x}' \in \mathcal{A}, \text{ and } \left\|\mathbf{x}' - \mathbf{x}\right\|^2 \leq \left\|\mathcal{P}_{\mathcal{A}}(\mathbf{x}) - \mathbf{x}\right\|^2 + \gamma, \tag{4.7}$$

so that $\mathcal{P}^{\gamma}_{\mathcal{A}}(\mathbf{x})$ yields a vector $\mathbf{x}' \in \mathcal{A}$ that approximately minimizes the squared distance from $\mathbf{x}$ to $\mathcal{A}$. Again, $\mathcal{P}^{\gamma}_{\mathcal{A}}(\mathbf{x})$ need not be uniquely defined for a particular input signal $\mathbf{x}$.

Certain specific instances of nonconvex manifolds do admit efficient exact projection operators. For example, consider the space of all K-sparse signals of length N; this can be viewed as the

Algorithm 5 Successive Projections onto INcoherent manifolds (SPIN)

Inputs: Observation matrix $\mathbf{\Phi}$, measurements $\mathbf{z}$, projection operators $\mathcal{P}_{\mathcal{A}}(\cdot), \mathcal{P}_{\mathcal{B}}(\cdot)$, number of iterations T, step size η
Outputs: Estimated signal components $\widehat{\mathbf{a}} \in \mathcal{A}, \widehat{\mathbf{b}} \in \mathcal{B}$
Initialize: $\mathbf{a}_0 = \mathbf{0}$, $\mathbf{b}_0 = \mathbf{0}$, $\mathbf{r} = \mathbf{z}$, $k = 0$
while $k \leq T$ **do**
  $\mathbf{g}_k \leftarrow \eta \mathbf{\Phi}^T \mathbf{r}$ {form gradient}
  $\widetilde{\mathbf{a}}_k \leftarrow \mathbf{a}_k + \mathbf{g}_k$, $\widetilde{\mathbf{b}}_k \leftarrow \mathbf{b}_k + \mathbf{g}_k$ {form signal proxies}
  $\mathbf{a}_{k+1} \leftarrow \mathcal{P}_{\mathcal{A}}(\widetilde{\mathbf{a}}_k)$, $\mathbf{b}_{k+1} \leftarrow \mathcal{P}_{\mathcal{B}}(\widetilde{\mathbf{b}}_k)$ {apply projection operators}
  $\mathbf{r} \leftarrow \mathbf{z} - \mathbf{\Phi}(\mathbf{a}_{k+1} + \mathbf{b}_{k+1})$ {update residual}
  $k \leftarrow k + 1$
end while
return $(\widehat{\mathbf{a}}, \widehat{\mathbf{b}}) \leftarrow (\mathbf{a}_T, \mathbf{b}_T)$

union of the $\binom{N}{K}$ canonical subspaces in $\mathbb{R}^N$ or, alternately, a K-dimensional submanifold of $\mathbb{R}^N$. Then, the projection of an arbitrary vector $\mathbf{x} \in \mathbb{R}^N$ onto this manifold is merely the best K-sparse approximation to $\mathbf{x}$, which can be very efficiently computed via simple thresholding. We discuss additional examples in Section 4.5.

4.4 The SPIN Algorithm

We now describe an algorithm to solve the linear inverse problem (4.1). Our proposed algorithm, *Successive Projections onto INcoherent manifolds* (SPIN), can be viewed as a generalization of several first-order methods for signal recovery for a variety of different models [48, 96, 123]. SPIN is described in pseudocode form in Algorithm 5.

The key innovation in SPIN is that we formulate *two* proxy vectors for the signal components $\widetilde{\mathbf{a}}_k$ and $\widetilde{\mathbf{b}}_k$ and project these onto the corresponding manifolds $\mathcal{A}$ and $\mathcal{B}$.

4.4.1 Recovery guarantees

We demonstrate that SPIN possesses strong *uniform recovery* guarantees comparable to existing state-of-the-art algorithms for sparse approximation and compressive sensing, while encompassing a very broad range of nonlinear signal models. The following theoretical result describes the performance of SPIN for signal recovery.

Theorem 2 (Main result)**.** *Suppose $\mathcal{A}, \mathcal{B}$ are ϵ-incoherent manifolds in $\mathbb{R}^N$. Let $\mathbf{\Phi}$ be a measurement matrix with restricted isometry constant δ over the direct sum manifold $\mathcal{C} = \mathcal{A} \oplus \mathcal{B}$. Suppose we observe noisy linear measurements $\mathbf{z} = \mathbf{\Phi}(\mathbf{a}^* + \mathbf{b}^*) + \mathbf{e}$, where $\mathbf{a}^* \in \mathcal{A}$ and $\mathbf{b}^* \in \mathcal{B}$. If*

$$0 \leq \delta < \frac{1 - 11\epsilon}{3 + 7\epsilon}, \tag{4.8}$$

then SPIN (Algorithm 5) with step size $\eta = 1/(1+\delta)$ with exact projections $\mathcal{P}_{\mathcal{A}}, \mathcal{P}_{\mathcal{B}}$ outputs $\mathbf{a}_T \in \mathcal{A}$

and $\mathbf{b}_T \in \mathcal{B}$*, such that* $\|\mathbf{z} - \mathbf{\Phi}(\mathbf{a}_T + \mathbf{b}_T)\|^2 \leq \beta \|\mathbf{e}\|^2 + \nu$ *in no more than* $T = \lceil \frac{1}{\log(1/\alpha)} \log \frac{\|\mathbf{z}\|^2}{2\nu} \rceil$ *iterations for any* $\nu > 0$*.*

Proof. See Appendix A. □

Here, $\alpha < 1$ and β are moderately-sized positive constants that depend only on δ and ϵ; we derive explicit expressions for α and β in Section A.1. For example, when $\epsilon = 0.05,\ \delta = 0.2$, we obtain $\alpha \approx 0.812,\ \beta \approx 5.404$.

For the special case when there is no measurement noise (i.e., $\mathbf{e} = 0$), Theorem 2 states that, after a finite number of iterations, SPIN outputs signal component estimates $(\widehat{\mathbf{a}}, \widehat{\mathbf{b}})$ such that $\left\|\mathbf{z} - \mathbf{\Phi}(\widehat{\mathbf{a}} + \widehat{\mathbf{b}})\right\| < \nu$ for any desired precision parameter ν. From the restricted isometry assumption on $\mathbf{\Phi}$ and Lemma 4, we immediately obtain Theorem 1. Since we can set ν to an arbitrarily small value, we have that the SPIN estimate $(\widehat{\mathbf{a}}, \widehat{\mathbf{b}})$ converges to the true signal pair $(\mathbf{a}^*, \mathbf{b}^*)$. Exact convergence of the algorithm might potentially take a very large number of iterations, but convergence to any desired positive precision constant β takes only a finite number of iterations. For the rest of the paper, we will informally denote signal "recovery" to imply convergence to a sufficiently fine precision.

SPIN assumes the availability of the exact projection operators $\mathcal{P}_{\mathcal{A}}, \mathcal{P}_{\mathcal{B}}$. In certain cases, it might be feasible to numerically compute only γ-approximate projections, as in (4.7). In this case, the bound on the norm of the error $\mathbf{z} - \mathbf{\Phi}(\mathbf{a}_T + \mathbf{b}_T)$ is only guaranteed to be upper bounded by a positive multiple of the approximation parameter γ. The following theoretical guarantee (with a near-identical proof mechanism as Theorem 2) captures this behavior.

Theorem 3 (Approximate projections)**.** *Under the same suppositions as Theorem 2, SPIN (Algorithm 5) with* γ*-approximate projections and step size* $\eta = 1/(1+\delta)$ *outputs* $\mathbf{a}_T \in \mathcal{A}$ *and* $\mathbf{b}_T \in B$ *such that* $\|\mathbf{z} - \mathbf{\Phi}(\mathbf{a}_T + \mathbf{b}_T)\|^2 \leq \beta \|\mathbf{e}\|^2 + \frac{1+\delta}{1-\alpha}\gamma + \nu$*, in no more than* $T = \lceil \frac{1}{\log(1/\alpha)} \log \frac{\|\mathbf{z}\|^2}{2\nu} \rceil$ *iterations.*

Proof. See Appendix A. □

We note some implications of Theorem 2. First, suppose that Φ is the identity operator, i.e., we have full measurements of the signal $\mathbf{x}^* = \mathbf{a}^* + \mathbf{b}^*$. Then, $\delta = 0$ and the lower bound on the restricted isometry constant holds with equality. However, we still require that $\epsilon < 1/11$ for guaranteed recovery using SPIN. We will discuss this condition further in Section 4.5.

Second, suppose that the one of the component manifolds is the trivial (zero) manifold; then, we have that $\epsilon = 0$. In this case, SPIN reduces to the Manifold Iterative Pursuit (MIP) algorithm for recovering signals from a single manifold [123]. Moreover, the condition on δ reduces to $0 \leq \delta < 1/3$, which exactly matches the condition required for guaranteed recovery using MIP.

Lastly, the condition (4.8) in Theorem 2 automatically implies that $\epsilon < 1/11$. This represents a mild tightening of the condition on ϵ required for a unique decomposition (Lemma 4), even with full measurements (i.e., when $\mathbf{\Phi}$ is the identity operator or, more generally, when $\delta = 0$).

4.5 Applications

The two-manifold signal model described in this paper is applicable to a wide variety of problems that have attracted considerable interest in the literature over the last several years. We discuss a

few representative instances and show how SPIN can be utilized for efficient signal recovery in each of these instances. We also present several numerical experiments that indicate the kind of gains that SPIN can offer in practice.

4.5.1 Sparse representations in pairs of bases

We revisit the classical problem of decomposing signals in an overcomplete dictionary that is the union of a pair of orthonormal bases and show how SPIN can be used to efficiently solve this problem. Let $\mathbf{\Psi}, \mathbf{\Psi'}$ be orthonormal bases of $\mathbb{R}^N$. Let $\mathcal{A}$ be the set of all K_1-sparse signals in $\mathbb{R}^N$ in the basis expansion of $\mathbf{\Psi}$, and let $\mathcal{B}$ be the set of all K_2- sparse signals in the basis expansion of $\mathbf{\Psi'}$. Then, $\mathcal{A}$ and $\mathcal{B}$ can be viewed as K_1- and K_2-dimensional submanifolds of $\mathbb{R}^N$, respectively. Consider a signal $\mathbf{x}^* = \mathbf{a}^* + \mathbf{b}^*$, where $\mathbf{a}^* \in \mathcal{A}, \mathbf{b}^* \in \mathcal{B}$ so that

$$\mathbf{a}^* = \sum_{i=1}^{K_1} a_i \boldsymbol{\psi}_i, \quad \mathbf{b}^* = \sum_{i=1}^{K_2} b_i \boldsymbol{\psi}'_i.$$

The problem is to recover $(\mathbf{a}^*, \mathbf{b}^*)$ given $\mathbf{x}^*$. This problem has been studied in many different forms in the literature, and several algorithms have been proposed in order to solve it efficiently [44, 119, 126]. See [127] for an in-depth study of the various state-of-the-art methods. All these methods assume a certain notion of incoherence between the two bases, most commonly referred to as the *mutual coherence* μ, which is defined as

$$\mu(\mathbf{\Psi}, \mathbf{\Psi'}) \triangleq \max_{i,j} \left| \langle \boldsymbol{\psi}_i, \boldsymbol{\psi}'_j \rangle \right|. \tag{4.9}$$

It is clear that $\mu \leq 1$. It can be shown that the mutual coherence also obeys the lower bound $\mu \geq 1/\sqrt{N}$ for any pair of bases of $\mathbb{R}^N$. This lower bound is in fact tight; equality is achieved, for example, when $\mathbf{\Psi}$ is the canonical basis in $\mathbb{R}^N$ and $\mathbf{\Psi'}$ is the basis defining the Walsh-Hadamard transform or the discrete Fourier basis [119, 126].

We establish the following simple relation between μ and the manifold incoherence between $\mathcal{A}$ and $\mathcal{B}$.

Lemma 6. *Let $\mathcal{A}$ be the set of all K_1-sparse signals in $\mathbf{\Psi}$, and $\mathcal{B}$ be the set of all K_2-sparse signals in $\mathbf{\Psi'}$. Let ϵ denote the manifold incoherence between $\mathcal{A}$ and $\mathcal{B}$. Then,*

$$\epsilon \leq \mu(\mathbf{\Psi}, \mathbf{\Psi'})(K_1 + K_2).$$

Proof. The secant manifold $\mathcal{S}(\mathcal{A})$ is equivalent to the set of signals in $\mathbb{R}^N$ that are $2K_1$-sparse in $\mathcal{A}$; similarly, $\mathcal{S}(\mathcal{B})$ is equivalent to the set of $2K_2$-sparse signals in $\mathcal{B}$. Therefore, if one considers unit norm vectors $\mathbf{u} \in \mathcal{S}(\mathcal{A}), \mathbf{u'} \in \mathcal{S}(\mathcal{B})$, we obtain

$$\begin{aligned} |\langle \mathbf{u}, \mathbf{u'} \rangle| &= \left| \left\langle \sum_{i=1}^{2K_1} a_i \lambda_i, \sum_{j=1}^{2K_2} b_j \lambda'_j \right\rangle \right| = \left| \sum_{i}^{2K_1} \sum_{j}^{2K_2} \langle \lambda_i, \lambda'_j \rangle a_i b_j \right| \\ &\leq \mu \sum_{i}^{2K_1} \sum_{j}^{2K_2} |a_i b_j| = \mu \left(\sum_{i=1}^{2K_1} |a_i| \right) \left(\sum_{j=1}^{2K_2} |b_j| \right), \end{aligned} \tag{4.10}$$

where the last relation follows from the triangle inequality. We can further bound the right hand side of (4.10). We have

$$\sum_{i=1}^{2K_1} |a_i| \leq \sqrt{2K_1} \sqrt{\sum_{i=1}^{2K_1} |a_i|^2} = \sqrt{2K_1} \|\mathbf{u}\| = \sqrt{2K_1},$$

since $\mathbf{u}$ is a unit vector. Similarly, $\sum_{i=1}^{2K_2} |b_i| \leq \sqrt{2K_2}$. Inserting these upper bounds in (4.10), we have

$$|\langle \mathbf{u}, \mathbf{u}' \rangle| \leq \mu \sqrt{2K_1} \sqrt{2K_2} \leq \mu(K_1 + K_2).$$

The lemma follows by considering the supremum over all vectors $\mathbf{u} \in \mathcal{S}(\mathcal{A}), \mathbf{u}' \in \mathcal{S}(\mathcal{B})$. □

We show how SPIN can be used to solve the linear inverse problem of recovering $(\mathbf{a}^*, \mathbf{b}^*)$ from $\mathbf{x}^*$. The restricted isometry assumption is not relevant in this case, since we assume that we have full measurements of the signal; therefore $\delta = 0$. An upper bound for the manifold incoherence parameter ϵ is specified in Lemma 6. The (exact) projection operators $\mathcal{P}_{\mathcal{A}}, \mathcal{P}_{\mathcal{B}}$ can be easily implemented; we simply perform a coefficient expansion in the corresponding orthornormal basis and retain the coefficients of largest magnitude. Mixing these ingredients together, we can guarantee that, given any signal $\mathbf{x}^*$, SPIN will return the true components $\mathbf{a}^*, \mathbf{b}^*$. This guarantee is summarized in the following result.

Corollary 1 (SPIN for pairs of bases)**.** *Let $\mathbf{x}^* = \mathbf{a}^* + \mathbf{b}^*$, where $\mathbf{a}^*$ is K_1-sparse in $\mathbf{\Psi}$ and $\mathbf{b}^*$ is K_2-sparse in $\mathbf{\Psi}'$. Let μ denote the mutual coherence between $\mathbf{\Psi}$ and $\mathbf{\Psi}'$. Then, SPIN exactly recovers $(\mathbf{a}^*, \mathbf{b}^*)$ from $\mathbf{x}^*$ provided*

$$K_1 + K_2 < \frac{1}{11\mu} \approx \frac{0.091}{\mu}. \tag{4.11}$$

Proof. If (4.11) holds, then from Lemma 6 we know that the manifold incoherence ϵ between $\mathcal{A}$ and $\mathcal{B}$ is smaller than 1/11. But this is exactly the condition required for guaranteed convergence of SPIN to the true signal components $(\mathbf{a}^*, \mathbf{b}^*)$. □

SPIN thus offers a conceptually simple method to separate mixtures of signals that are sparse in incoherent bases. The only condition required on the signals is that the total sparsity $K_1 + K_2$ is upper-bounded by the quantity $0.09/\mu$. The best known approach for this problem is an ℓ_1-minimization formulation that features a similar guarantee that is known to be tight [126, 128]:

$$K_1 + K_2 < \frac{\sqrt{2} - 0.5}{\mu} \approx \frac{0.914}{\mu}.$$

Therefore, SPIN yields a recovery guarantee that is off the best possible method by a factor of 10. Once again, it is possible that the constant 1/11 in Corollary 1 can be tightened by a more careful analysis of SPIN specialized to the case when the component signal manifolds correspond to a pair of incoherent bases, but we will not pursue this direction here. It is also possible to generalize SPIN to the case where the sparsifying dictionary comprises a union of more than two orthonormal bases [129]; see Section 4.6 for a short discussion.

4.5.2 Articulation manifolds

Articulation manifolds provide a powerful, flexible conceptual tool for modeling signals and image ensembles in a number of applications [28, 130]. Consider an ensemble of signals $\mathcal{M} \subset \mathbb{R}^N$ that are generated by varying K parameters $\boldsymbol{\theta} \in \Theta,\ \Theta \subset \mathbb{R}^K$. Then, we say that the signals trace out a nonlinear K-dimensional articulation manifold in $\mathbb{R}^N$, where $\boldsymbol{\theta}$ is called the *articulation parameter vector*. Examples of articulation manifolds include: acoustic chirp signals of varying frequencies (where $\boldsymbol{\theta}$ represents the chirp rate); images of a white disk translating on a black background (where $\boldsymbol{\theta}$ represents the planar location of the disk center); and images of a solid object with variable pose (where $\boldsymbol{\theta}$ represents the six-dimensional pose parameters, three corresponding to spatial location and three corresponding to orientation).

We consider the class of *compact, smooth*, K-dimensional articulation manifolds $\mathcal{M} \subset \mathbb{R}^N$. For such manifold classes, it is possible to construct linear measurement operators $\boldsymbol{\Phi}$ that preserve the pairwise secant geometry of $\mathcal{M}$. Specifically, it has been shown [82] that there exist *randomized* constructions of measurement operators $\boldsymbol{\Phi} \in \mathbb{R}^{M \times N}$ that satisfy the RIP on the secants of $\mathcal{M}$ with constant δ and with probability at least ρ, provided

$$M = \mathcal{O}\left(K \frac{\log(C_{\mathcal{M}} N \delta^{-1}) \log(\rho^{-1})}{\delta^2}\right)$$

for some constant C that depends only on the smoothness and volume of the manifold $\mathcal{M}$. Therefore, the dimension of the range space of $\boldsymbol{\Phi}$ is proportional to the *number of degrees of freedom* K, but is only logarithmic in the ambient dimension N. Moreover, given such a measurement matrix $\boldsymbol{\Phi}$ with isometry constant $\delta < 1/3$ and a projection operator $\mathcal{P}_{\mathcal{M}}(\cdot)$ onto $\mathcal{M}$, any signal $\mathbf{x} \in \mathcal{M}$ can be reconstructed from its compressive measurements $\mathbf{y} = \boldsymbol{\Phi}\mathbf{x}$ using Manifold Iterative Pursuit (MIP) [123].

We generalize this setting to the case where the unknown signal of interest arises as a mixture of signals from two manifolds $\mathcal{A}$ and $\mathcal{B}$. For instance, suppose we are interested in the space of images, where $\mathcal{A}$ and $\mathcal{B}$ comprise of translations of fixed template images $f(\mathbf{t})$ and $g(\mathbf{t})$, where $\mathbf{t}$ denotes the 2D domain over which the image is defined. Then, the signal of interest is an image of the form

$$\mathbf{x}^* = \mathbf{a}^* + \mathbf{b}^* = f(\mathbf{t} + \boldsymbol{\theta}_1) + g(\mathbf{t} + \boldsymbol{\theta}_2),$$

where $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ denote the unknown translation parameters. The problem is to recover $(\mathbf{a}^*, \mathbf{b}^*)$, or equivalently $(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$, given compressive measurements $\mathbf{z} = \boldsymbol{\Phi}(\mathbf{a}^* + \mathbf{b}^*)$.

We demonstrate that SPIN offers an easy, efficient technique to recover the component images. This example also demonstrates that SPIN is robust to practical considerations such as noise. Figure 4.1 displays the results of SPIN recovery of a 64×64 image from very limited measurements. The unknown image consists of the linear sum of arbitrary translations of template images $f(\mathbf{t})$ and $g(\mathbf{t})$, that are smoothed binary images on a black background of a white disk and a white square, respectively. Further, the image has been contaminated with significant Gaussian noise (SNR = 14dB) prior to measurement (Fig. 4.1(a)). From Figs. 4.1(b) and 4.1(c), we observe that SPIN is able to perfectly recover the original component signals from merely $M = 50$ random linear measurements.

(a) Original image (b) Disk component (c) Square component

Figure 4.1: *SPIN recovery of a noisy* 64×64 *image from compressive measurements. The clean image consists of the linear superposition of a disk and a square of fixed pre-specified sizes, but the locations of the centers of the disk and the square are unknown. Additive Gaussian noise (SNR = 14dB) is added to the image prior to measurement. Signal length* $N = 64 \times 64 = 4096$*, number of compressive measurements* $M = 50$*. (a) Original noisy image. (b) Reconstructed disk. (c) Reconstructed square. SPIN perfectly reconstructs both components from just* $M/N = 1.2\%$ *measurements.*

For guaranteed SPIN convergence, we require that the manifolds $\mathcal{A}, \mathcal{B}$ are incoherent. Informally, the condition of incoherence on the secants of $\mathcal{A}$ and B is always valid when the template images $f(\mathbf{t}), g(\mathbf{t})$ are "sufficiently" distinct. This intuition is made precise using the bound in (4.3). More generally, we can state the following theoretical guarantee for SPIN performance in the case of general higher-dimensional manifolds.

Corollary 2 (SPIN for pairs of manifolds)**.** *Let the entries of* $\mathbf{\Phi} \in \mathbb{R}^{M \times N}$ *be chosen from a standard Gaussian probability distribution. Let* $\mathcal{A}$ *and* $\mathcal{B}$ *be* ϵ*-incoherent compact submanifolds of* $\mathbb{R}^N$ *of dimensions* K *and* K' *respectively. Let* $\mathbf{z} = \mathbf{\Phi}(\mathbf{a}^* + \mathbf{b}^*)$*, where* $\mathbf{a}^* \in \mathcal{A}$ *and* $\mathbf{b}^* \in \mathcal{B}$*. Then, with high probability, SPIN exactly recovers* $(\mathbf{a}^*, \mathbf{b}^*)$ *from* $\mathbf{z}$*, provided*

$$M = \mathcal{O}\left(\left(K \log(C_{\mathcal{A}} N) + K' \log(C_{\mathcal{B}} N)\right)\right). \tag{4.12}$$

Here, $C_{\mathcal{A}}, C_{\mathcal{B}}$ *are constants that depend only on certain intrinsic geometric parameters (such as the volume) of* $\mathcal{A}, \mathcal{B}$ *respectively.*

Proof. It is easy to see that if the matrix $\mathbf{\Phi}$ satisfies the RIP on the secants of the direct sum $\mathcal{C} = \mathcal{A} \oplus \mathcal{B}$, then SPIN recovery follows from Theorem 2. We show that a randomized construction of $\mathbf{\Phi}$ with number of rows specified by (4.12) satisfies the RIP on $\mathcal{C}$ with high probability. Essentially, our proof combines the techniques used in Section 3.2 of [82] with Lemma 1 of [131].

The manifold-embedding result in [82] is proved using two fundamental steps: (i) careful construction of a *finite* subset $\mathcal{R}$ of points in $\mathbb{R}^N$ that serves as a dense covering of the K-dimensional manifold of interest $\mathcal{A}$; and (ii) application of the Johnson-Lindenstrauss lemma [81] to this finite set $\mathcal{R}$ to produce, with vanishingly low probability of failure, a Gaussian measurement matrix $\mathbf{\Phi}$ that satisfies the RIP on $\mathcal{A}$. Section 3.2.5 of [82] indicates that the cardinality of the the finite set $\mathcal{R}$ can be upper bounded as

$$\#\mathcal{R} \leq (C_{\mathcal{A}} N)^K,$$

where $C_{\mathcal{A}}$ is a constant that depends only on the intrinsic geometry of $\mathcal{A}$. However, in our setting we are interested in the direct sum of manifolds; correspondingly, we can construct finite sets $\mathcal{R}_{\mathcal{A}}$ and $\mathcal{R}_{\mathcal{B}}$ and apply Lemma 1 of [131], that specifies a lower bound on the number of measurements required to preserve the norms of linear sums of finite point sets:

$$M \geq \mathcal{O}\left(\log(\mathcal{R}_{\mathcal{A}}\mathcal{R}_{\mathcal{B}})\right) = \mathcal{O}\left(\left(K\log(C_{\mathcal{A}}N) + K'\log(C_{\mathcal{B}}N)\right)\right),$$

where K, K' are the dimensions of $\mathcal{A}, \mathcal{B}$ respectively. Corollary 2 follows. □

An important consideration in SPIN is the tractable computation of the projection $\mathcal{P}_{\mathcal{M}}(\mathbf{x})$ given any $\mathbf{x} \in \mathbb{R}^N$. For example, in the numerical example in Fig. 4.1, the operator $\mathcal{P}_{\mathcal{A}}(\mathbf{x})$ onto the manifold $\mathcal{A}$ consists of running a matched filter between the template $f(\mathbf{t})$ and the input signal $\mathbf{x}$ and returning $f(\mathbf{t}+\widehat{\boldsymbol{\theta}})$, where the parameter value $\widehat{\boldsymbol{\theta}}$ corresponds to the 2D location of the maximum of the matched filter response. This is very efficiently carried out in $\mathcal{O}\left(N\log N\right)$ operations using the Fast Fourier Transform (FFT). However, for more complex articulation manifolds, the exact projection operation might not be tractable, whereas only approximate numerical projections can be efficiently computed. In this case also, SPIN can recover the signal components $(\mathbf{a}^*, \mathbf{b}^*)$, but with weaker convergence guarantees (Theorem 3).

4.5.3 Signals in impulsive noise

In some situations, the signal of interest $\mathbf{x}$ might be corrupted with *impulsive noise* (or shot noise) prior to signal acquisition via linear measurements. For example, consider Fig. 4.2(a), where the Gaussian pulse is the signal of interest, and the spikes indicate the undesirable noise. In this case, the linear observations are more accurately modeled as:

$$\mathbf{z} = \boldsymbol{\Phi}(\mathbf{x} + \mathbf{n}), \quad \text{such that } \mathbf{x} \in \mathcal{M},$$

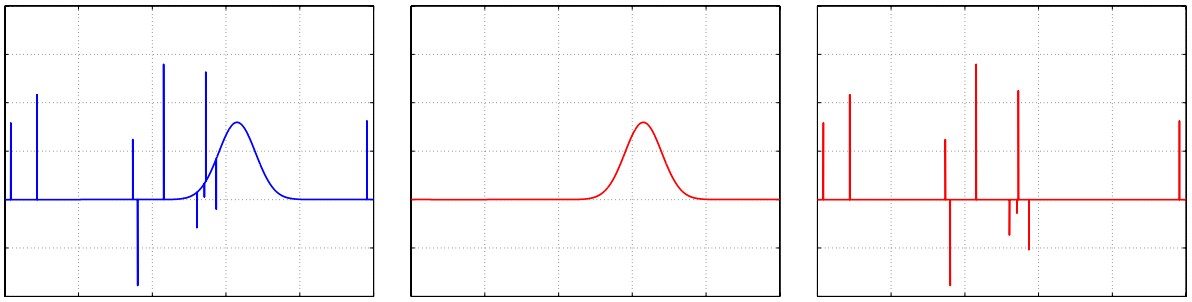
and $\mathbf{n}$ is a K'-sparse signal in the canonical basis. Therefore, SPIN can be used to recover $\mathbf{x}$ from $\mathbf{z}$, provided that the manifold $\mathcal{M}$ is incoherent with the set of sparse signals $\Sigma_{K'}$ and $\boldsymbol{\Phi}$ satisfies the RIP on the direct sum $\mathcal{M} + \Sigma_{K'}$. We summarize this result in the form of the following Lemma.

Lemma 7 (SPIN for robust recovery)**.** *Let the entries of $\boldsymbol{\Phi} \in \mathbb{R}^{M\times N}$ be independently chosen from a standard normal probability distribution. Let $\mathcal{M} \subset \mathbb{R}^N$ be a $K-$dimensional articulation manifold, and $\Sigma_{K'}$ be the set of all K'-sparse signals in $\mathbb{R}^N$. Then, with high probability, $\boldsymbol{\Phi}$ satisfies the restricted isometry property (RIP) with constant $\delta \in [0,1]$ on secants of the direct sum $\mathcal{C} = \mathcal{M}+\Sigma_{K'}$, provided:*

$$M \geq (C_1 K + C_2 K')\frac{\log(C_3 N)}{\delta^2},$$

where C_1, C_2, C_3 are universal constants.

Figure 4.2 displays the results of a numerical experiment that illustrates the utility of SPIN in this setting. We consider a manifold of signals of length $N = 10000$ that consist of shifts of a Gaussian pulse of fixed width $\mathbf{g}_0 \in \mathbb{R}^N$. The unknown signal $\mathbf{x}$ is an element of this manifold $\mathcal{M}$, and is corrupted by $K' = 10$ spikes of unknown magnitudes and locations. This degraded signal is sampled using $M = 150$ random linear measurements to obtain an observation vector $\mathbf{z}$.



(a) Original signal (b) Manifold component (c) Noise component

Figure 4.2: *SPIN recovery of a shifted Gaussian pulse from compressive measurements. The shift parameter of the pulse is unknown, and the signal is corrupted with K'-sparse, impulsive noise of unknown amplitudes and locations. $N = 10000, K' = 10, M = 150$. (a) Original signal. (b) Reconstructed Gaussian pulse (Recovery SNR = 80.09 dB). (c) Estimated noise component. SPIN perfectly reconstructs both components from just $M/N = 1.5\%$ measurements.*

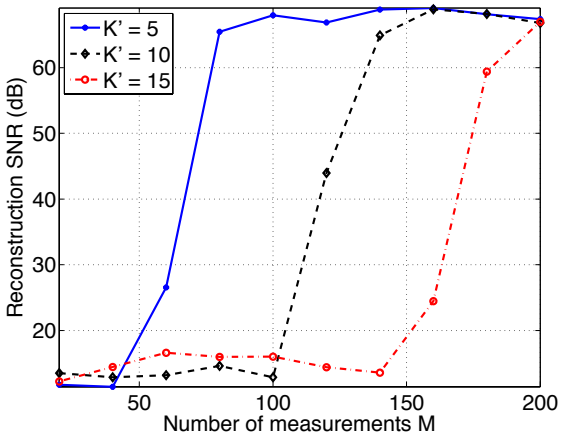


Figure 4.3: *Monte Carlo simulation of SPIN signal recovery in impulsive noise , averaged over 100 trials. In each trial, the measured signal is the sum of a randomly shifted Gaussian pulse and a random K'-sparse signal. SPIN can tolerate a higher number K' nuisance impulses by increasing the number of measurements M. By Corollary 2 this dependence of M on K' can be shown to be* linear.

We apply SPIN to recover $\mathbf{x}$ from $\mathbf{z}$. The projection operator $\mathcal{P}_{\mathcal{M}}(\cdot)$ consists of a matched filter with the template pulse $\mathbf{g}_0$, while the projection operator $\mathcal{P}_{\Sigma_{K'}}(\cdot)$ simply returns the best K'-term approximation in the canonical basis. Assuming that we have knowledge of the number of nonzeros in the noise vector $\mathbf{n}$, we can use SPIN to reconstruct both $\mathbf{x}$ and $\mathbf{n}$. We observe from Fig. 4.2(b) that SPIN recovers the true signal $\mathbf{x}$ with near-perfect accuracy. Further, this recovery is possible with only a small number $M = 150$ linear measurements of $\mathbf{x}$, which constitutes but a fraction of the ambient dimension of the signal space.

Figure 4.3 plots the number of measurements M vs. the signal reconstruction error (normalized relative to the signal energy and plotted in dB). We observe that, by increasing M, SPIN can tolerate an increased number K' of nuisance spikes. Further, by Corollary 2, we observe that this relationship between M and K' is in fact *linear*. This result can be extended to any situation where the signals of interest obey a "hybrid" model that is a mixture of a nonlinear manifold and the set of sparse signals.

4.6 Discussion

We have proposed and rigorously analyzed an algorithm, which we dub Successive Projections onto INcoherent Manifolds (SPIN), for the recovery of a pair of signals given a small number of measurements of their linear sum. For SPIN to guarantee signal recovery, we require two main geometric criteria to hold: (i) the component signals should arise from two disjoint manifolds that are in a specific sense *incoherent*, and (ii) the linear measurement operator should satisfy a *restricted isometry* criterion on the secants of the direct sum of the two manifolds. The computational efficiency of SPIN is determined by the tractability of the *projection operators* onto either component manifold. We have presented indicative numerical experiments demonstrating the utility of SPIN, but defer a thorough experimental study of SPIN to future work.

SPIN is an iterative gradient projection algorithm and requires as input parameters the number of iterations T and the gradient step size η. The iteration count T can be chosen using one of many commonly-used stopping criteria. For example, convergence can be declared if the norm of the error $\psi(\mathbf{a}_k, \mathbf{b}_k)$ at the $(k+1)$-th time step does not differ significantly from the error at k-th time step. The choice of optimal step size η is more delicate. Theorem 2 relates the step size to the restricted isometry constant δ of $\mathbf{\Phi}$, but this constant is not easy to calculate. In our preliminary findings, a step size in the range $0.5 \leq \eta \leq 0.7$ consistently gave good results. See [132] for a discussion on the choice of step size for hard thresholding methods.

In practical scenarios, the signal of interest rarely belongs exactly to a low-dimensional submanifold $\mathcal{M}$ of the ambient space, but is only well-approximated by $\mathcal{M}$. Interestingly, in such situations the effect of this mismatch can be studied using the concept of γ-approximate projections (4.7). Theorem 3 rigorously demonstrates that SPIN is robust to such approximations. Further, our main result (Theorem 2) indicates that SPIN is stable with respect to inaccurate measurements, owing to the fact that the reconstruction error is bounded by a constant times the norm of the measurement noise vector $\mathbf{e}$.

For clarity and brevity, we have focused our attention on signals belonging to the direct sum of two signal manifolds. However, SPIN (and its accompanying proof mechanism) can be conceptually extended to sums of any Q manifolds. In such a scenario, the conditions of convergence of SPIN would require that the component manifolds are Q-wise incoherent, and the measurement operator $\mathbf{\Phi}$ satisfies a restricted isometry on the Q-wise direct sum of the component manifolds.

An intriguing open question is whether SPIN (or a similar first-order projected gradient algorithm) is applicable to situations where either of the component manifolds is the set of low-rank matrices. The problem of reconstructing, from affine measurements, matrices that are a sum of low-rank and sparse matrices has attracted significant attention in the recent literature [59, 120, 133].

The key stumbling block is that the manifold of low-rank matrices is *not incoherent* with the manifold of sparse matrices; indeed, the two manifolds share a nontrivial intersection (i.e., there exist low rank matrices that are also sparse, and vice versa). We revisit this problem in Chapter 7.

CHAPTER 5

Random Projections for Manifold Learning

5.1 Setup

In this chapter, we consider the problem of efficient nonlinear dimensionality reduction, or *manifold learning*, of high-dimensional signal ensembles. Most conventional manifold learning algorithms are adaptive (i.e., data dependent) and nonlinear (i.e., involve construction of a nonlinear mapping into a smaller space). However, a *linear, nonadaptive* manifold dimensionality reduction technique has emerged that employs the technique of *random projections* [82]. Consider a K-dimensional manifold $\mathcal{M}$ in the ambient space $\mathbb{R}^N$ and its projection onto a random subspace of dimension $M = CK\log(N)$; note that $K < M \ll N$. The central result of [82] is that the pairwise metric structure of sample points from $\mathcal{M}$ is preserved with high accuracy under a random linear projection.

This result has far reaching implications. Prototypical devices that directly and inexpensively acquire random projections of certain types of data (signals, images, etc.) have been developed [136, 137]; these devices are hardware realizations of the mathematical tools developed in the emerging area of Compressed Sensing (CS) [19, 138]. The advantages of random projections extend even to cases where the original data is available in the ambient space $\mathbb{R}^N$. For example, consider a wireless network of cameras observing a scene. To perform joint image analysis, the following steps might be executed:

1. **Collate**: Each camera node transmits its respective captured image (of size N) to a central processing unit.

2. **Preprocess**: The central processor estimates the *intrinsic dimension* K of the underlying image manifold.

This work is in collaboration with Richard G. Baraniuk, Mark A. Davenport, Marco F. Duarte, and Michael B. Wakin [134, 135].

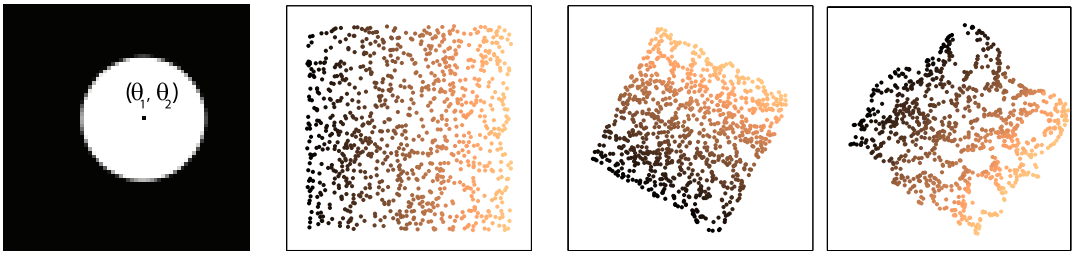


Figure 5.1: *(a) Input data consisting of 1000 images of a disk shifted in $K = 2$ dimensions, parametrized by an articulation vector $(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$. (b) True $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ values of the sampled data. (c) Isomap embedding learned from original data in $\mathbb{R}^N$. (d) Isomap embedding learned from merely $M = 15$ random projections.*

3. **Learn**: The central processor performs a nonlinear embedding of the data points – for instance, using Isomap (described in Algorithm 2) – into a K-dimensional Euclidean space, using the estimate of K from the previous step.

In situations where N is large and communication bandwidth is limited, the dominating costs will be in the first transmission/collation step. On the one hand, to reduce the communication needs one may perform nonlinear image compression (such as JPEG) at each node before transmitting to the central processing. But this requires a good deal of processing power at each sensor, and the compression would have to be undone during the learning step, thus adding to overall computational costs. On the other hand, every camera could encode its image by computing (either directly or indirectly) a small number of random projections to communicate to the central processor. These random projections are obtained by linear operations on the data, and thus are cheaply computed. Clearly, in many situations it will be less expensive to store, transmit, and process such randomly projected versions of the sensed images. In such situations, the following question is paramount: how much information about the manifold is conveyed by these random projections, and is any advantage in analyzing such measurements from a manifold learning perspective?

In this chapter, we develop and analyze *efficient* methods for manifold learning of high-dimensional data ensembles. We provide theoretical and experimental evidence that via the technique of random projections, reliable learning of a low-dimensional manifold can be performed not just in the high-dimensional ambient space $\mathbb{R}^N$ but also in an intermediate, much lower-dimensional random projection space.

We present a theoretical bound on the minimum number of measurements per sample point required to estimate the intrinsic dimension (ID) of the underlying manifold, up to an accuracy level comparable to that of the Grassberger-Procaccia algorithm [139, 140], a widely used geometric approach for dimensionality estimation. We present a similar bound on the number of measurements required for Isomap [60] – a popular manifold learning algorithm – to be "reliably" used to discover the nonlinear structure of the manifold. Additionally, we formulate a procedure to determine, in practical settings, this minimum value of M with no *a priori* information about the data points. See, for example, the toy example of Figure 5.1 for an illustration of this approach.

We also extend our analysis to multi-sensor scenarios where a collection of signal ensembles are governed by a small number of common dependency parameters. We propose a new *joint manifold*

model that captured such dependencies. In conjunction with our proposed random projections-based manifold learning approach, we can formulate a *scalable* dimensionality reduction scheme that efficiently fuses the data from all sensors. Our analysis is particularly relevant in distributed sensing systems and leads to significant potential savings in data acquisition, storage and transmission costs. We demonstrate these gains via a number of experiments using real-world image data.

The rest of this chapter is organized as follows. In Section 5.2 we briefly describe the manifold learning approaches that we consider. In Section 5.3 we present our main theoretical contributions, namely, the bounds on M required to perform reliable dimensionality estimation and manifold learning from random projections. We also describes a new adaptive algorithm that estimates the minimum value of M required to provide a faithful representation of the data so that manifold learning can be performed. In Section 5.5 we introduce the joint manifold model for multi-sensor signal ensembles, and apply this model to develop an efficient data fusion scheme based on the random projections approach. Experimental results on a variety of real-world datasets are provided in Section 5.6. In Section 5.7 we provide a concluding summary.

5.2 Related Work

An important input parameter for all manifold learning algorithms is the *intrinsic dimension* (ID) of a point cloud. We aim to embed the data points in as low-dimensional a space as possible in order to avoid the curse of dimensionality. However, if the embedding dimension is too small, then distinct data points might be collapsed onto the same embedded point. Hence a natural question to ask is: given a point cloud in N-dimensional Euclidean space, what is the dimension of the manifold that best captures the structure of this data set? This problem has received considerable attention in the literature and remains an active area of research[139, 141, 142].

For the purposes of this paper, we focus our attention on the Grassberger-Procaccia (GP) [139] algorithm for ID estimation. This is a widely used geometric technique that takes as input the set of pairwise distances between sample points. It then computes the *scale-dependent correlation dimension* of the data, defined as follows.

Definition 6. *Suppose $\mathcal{X} = (\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n)$ is a finite dataset of underlying dimension K. Define*

$$C_n(r) = \frac{1}{n(n-1)} \sum_{i \neq j} \mathbf{I}_{\|\mathbf{x}_i - \mathbf{x}_j\| < r},$$

where $\mathbf{I}$ is the indicator function. The scale-dependent correlation dimension of $\mathcal{X}$ is defined as

$$\widehat{D}_{\text{corr}}(r_1, r_2) = \frac{\log C_n(r_1) - \log C_n(r_2)}{\log r_1 - \log r_2}.$$

The best possible approximation to K (call this $\widehat{K}$) is obtained by fixing r_1 and r_2 to the biggest range over which the plot is linear and the calculating D_{corr} in that range. There are a number of practical issues involved with this approach; indeed, it has been shown that geometric ID estimation algorithms based on finite sampling yield biased estimates of intrinsic dimension [142, 143]. In our theoretical derivations, we do not attempt to take into account this bias; instead, we prove that

the effect of running the GP algorithm on a sufficient number of random projections produces a dimension estimate that well-approximates the GP estimate obtained from analyzing the original point cloud.

The estimate $\widehat{K}$ of the ID of the point cloud is used by nonlinear manifold learning algorithms (e.g., Isomap [60], Locally Linear Embedding (LLE) [61], and Hessian Eigenmaps [72], among many others) to generate a $\widehat{K}$-dimensional coordinate representation of the input data points. Our main analysis will be centered around Isomap (detailed in Algorithm 2). Isomap attempts to preserve the *metric structure* of the manifold, i.e., the set of pairwise geodesic distances of any given point cloud sampled from the manifold. In essence, Isomap approximates the geodesic distances using a suitably defined graph and performs classical multidimensional scaling (MDS) to obtain a reduced K-dimensional representation of the data [60].

A key parameter in the Isomap algorithm is the *residual variance*, which is equivalent to the stress function encountered in classical MDS. The residual variance is a measure of how well the given dataset can be embedded into a Euclidean space of dimension K. In the next section, we prescribe a specific number of measurements per data point so that performing Isomap on the randomly projected data yields a residual variance that is arbitrarily close to the variance produced by Isomap on the original dataset.

We conclude this section by revisiting the results derived in [82], which form the basis for our development. Consider the effect of projecting a smooth K-dimensional manifold residing in $\mathbb{R}^N$ onto a random M-dimensional subspace (isomorphic to $\mathbb{R}^M$). If M is sufficiently large, a stable near-isometric embedding of the manifold in the lower-dimensional subspace is ensured. The key advantage is that M needs only to be *linear in the intrinsic dimension* of the manifold K. In addition, M depends only logarithmically on intrinsic geometric properties of the manifold, such as its volume and condition number. The result can be summarized in the following theorem.

Theorem 4. [82] *Let $\mathcal{M}$ be a compact K-dimensional manifold in $\mathbb{R}^N$ having volume V and condition number $1/\tau$. Fix $0 < \delta < 1$ and $0 < \rho < 1$. Let $\mathbf{\Phi}$ be a random* orthoprojector*, formed by orthogonalizing M vectors of length N having i.i.d. Gaussian or Bernoulli distributed entries, from $\mathbb{R}^N$ to $\mathbb{R}^M$. Suppose that*

$$M \geq O\left(\frac{K \log(NV\tau^{-1}) \log(\rho^{-1})}{\delta^2}\right). \tag{5.1}$$

Also, suppose $M < N$. Then, with probability exceeding $1 - \rho$, the following statement holds: For every pair of points $\mathbf{x}, \mathbf{y} \in \mathcal{M}$, and $i \in \{1, 2\}$,

$$(1 - \delta)\sqrt{\frac{M}{N}} \leq \frac{d_i(\mathbf{\Phi x}, \mathbf{\Phi y})}{d_i(\mathbf{x}, \mathbf{y})} \leq (1 + \delta)\sqrt{\frac{M}{N}}. \tag{5.2}$$

where $d_1(\mathbf{x}, \mathbf{y})$ (respectively, $d_2(\mathbf{x}, \mathbf{y})$) stands for the geodesic (respectively, ℓ_2) distance between points $\mathbf{x}$ and $\mathbf{y}$.

Note that (5.2) is again similar to the criteria of *restricted isometry* introduced earlier in (2.8) and (3.1), except that here the notion of restricted isometry extends over all points $\mathbf{x}, \mathbf{y}$ belonging to $\mathcal{M}$. Once again, we refer to the scalar quantity δ as the "isometry constant".

The condition number τ controls the local, as well as global, curvature of the manifold – the smaller the τ, the less well-conditioned the manifold with higher "twistedness" [82]. Theorem 4 has been proved by first specifying a finite high-resolution sampling on the manifold, the nature of which depends on its intrinsic properties; for instance, a planar manifold can be sampled coarsely. Then the Johnson-Lindenstrauss Lemma [88] is applied to these points to guarantee the isometry constant δ.

5.3 Learning with Random Projections: Theory

We saw above that random projections essentially ensure that the metric structure of a high-dimensional input point cloud (i.e., the set of all pairwise distances between points belonging to the dataset) is preserved up to a distortion that depends on δ. This immediately suggests that geometry-based ID estimation and manifold learning algorithms could be applied to the lower-dimensional, randomly projected version of the dataset. In this Section, we rigorously justify this intuition.

5.3.1 The GP algorithm

The first of our main theoretical results establishes a sufficient dimension of random projection M required to maintain the fidelity of the estimated correlation dimension using the GP algorithm. The proof of the following is detailed in [144].

Theorem 5. *Let $\mathcal{M}$ be a compact K-dimensional manifold in $\mathbb{R}^N$ having volume V and condition number $1/\tau$. Let $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, ...\}$ be a sequence of samples drawn from a* uniform *density supported on $\mathcal{M}$. Let $\widehat{K}$ be the dimension estimate of the GP algorithm on $\mathcal{X}$ over the range $(r_{\min}, r_{\max})$. Let $\beta = \ln(r_{\max}/r_{\min})$. Fix $0 < \delta' < 1$ and $0 < \rho < 1$. Suppose that $r_{\max} < \tau/2$. Let $\mathbf{\Phi}$ be a random orthoprojector from $\mathbb{R}^N$ to $\mathbb{R}^M$ with $M < N$ and*

$$M \geq O\left(\frac{K \log(NV\tau^{-1}) \log(\rho^{-1})}{\beta^2 \delta'^2}\right). \tag{5.3}$$

Let $\widehat{K}_{\mathbf{\Phi}}$ be the estimated correlation dimension on $\mathbf{\Phi}\mathcal{X}$ in the projected space *over the range $(r_{min}\sqrt{M/N}, r_{max}\sqrt{M/N})$. Then, $\widehat{K}_{\mathbf{\Phi}}$ is bounded by:*

$$(1 - \delta')\widehat{K} \leq \widehat{K}_{\mathbf{\Phi}} \leq (1 + \delta')\widehat{K} \tag{5.4}$$

with probability exceeding $1 - \rho$.

Proof. See Appendix B. □

Theorem 5 is a worst-case bound and serves as a sufficient condition for stable ID estimation using random projections. Thus, if we choose a sufficiently small value for δ' and ρ, we are guaranteed estimation accuracy levels as close as desired to those obtained with ID estimation in the original signal space. Note that the bound on $\widehat{K}_{\mathbf{\Phi}}$ is *multiplicative*. This implies that in the worst case, the number of projections required to estimate $\widehat{K}_{\mathbf{\Phi}}$ very close to $\widehat{K}$ (say, within integer roundoff error) becomes higher with increasing manifold dimension K.

5.3.2 Isomap

The second of our main results prescribes the minimum dimension of random projections required to maintain the residual variance produced by Isomap in the projected domain within an arbitrary *additive* constant of that produced by Isomap with the full data in the ambient space. This proof of this theorem [144] relies on the proof technique used in [76].

Theorem 6. *Let $\mathcal{M}$ be a compact K-dimensional manifold in $\mathbb{R}^N$ having volume V and condition number $1/\tau$. Let $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n\}$ be a finite set of samples drawn from a sufficiently fine density supported on $\mathcal{M}$. Let $\mathbf{\Phi}$ be a random orthoprojector from $\mathbb{R}^N$ to $\mathbb{R}^M$ with $M < N$. Fix $0 < \delta < 1$ and $0 < \rho < 1$. Suppose*

$$M \geq O\left(\frac{K \log(NV\tau^{-1}) \log(\rho^{-1})}{\delta^2}\right).$$

Define the diameter Γ *of the dataset as follows:*

$$\Gamma = \max_{1 \leq i,j \leq n} d_{\text{iso}}(\mathbf{x}_i, \mathbf{x}_j)$$

where $d_{\text{iso}}(\mathbf{x}, \mathbf{y})$ stands for the Isomap estimate of the geodesic distance between points $\mathbf{x}$ and $\mathbf{y}$. Define R and $R_{\mathbf{\Phi}}$ to be the residual variances obtained when Isomap generates a K-dimensional embedding of the original dataset $\mathcal{X}$ and projected dataset $\mathbf{\Phi}\mathcal{X}$ respectively. Under suitable constructions of the Isomap connectivity graphs, $R_{\mathbf{\Phi}}$ is bounded by:

$$R_{\mathbf{\Phi}} < R + C\Gamma^2\delta,$$

with probability exceeding $1 - \rho$. C is a function only on the number of sample points n.

Proof. See Appendix B. □

Since the choice of δ is arbitrary, we can choose a large enough M (which is still only logarithmic in N) such that the residual variance yielded by Isomap on the randomly projected version of the dataset is arbitrarily close to the variance produced with the data in the ambient space. Again, this result is derived from a worst-case analysis. Note that Γ acts as a measure of the scale of the dataset. In practice, we may enforce the condition that the data is normalized (i.e., every pairwise distance calculated by Isomap is divided by Γ). This ensures that the K-dimensional embedded representation is contained within a ball of unit norm centered at the origin.

5.4 Learning with Random Projections: Practice

Thus, we have proved that with only an M-dimensional projection of the data (with $M \ll N$) we can perform ID estimation and subsequently learn the structure of a K-dimensional manifold, up to accuracy levels obtained by conventional methods. Note that this is only a sufficiency result, and the actual gains can be considerably higher. Further, in practice, it is hard to know or estimate the parameters V and τ of the underlying manifold. Also, since we have no *a priori* information regarding the data, it is impossible to fix $\widehat{K}$ and R, the outputs of GP and Isomap on the point

Algorithm 6 ML-RP

$M \leftarrow 1$
$\mathbf{\Phi} \leftarrow$ Random orthoprojector of size $M \times N$.
while residual variance $\geq \gamma$ **do**
  Run the GP algorithm on $\mathbf{\Phi}\mathcal{X}$.
  Use ID estimate $(\widehat{K})$ to perform Isomap on $\mathbf{\Phi}\mathcal{X}$.
  Calculate residual variance.
  $M \leftarrow M + 1$
  Add one row to $\mathbf{\Phi}$
end while
return M
return $\widehat{K}$

cloud in the ambient space. Thus, often, we may not be able fix a definitive value for M. To circumvent this problem we develop the following greedy procedure that we dub it **ML-RP** for *manifold learning using random projections*.

We initialize M to a small number, and compute M random projections of the data set $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n\}$ (here n denotes the number of points in the point cloud). Using the set $\mathbf{\Phi}\mathcal{X} = \{\mathbf{\Phi}\mathbf{x} : \mathbf{x} \in \mathcal{X}\}$, we estimate the intrinsic dimension using the GP algorithm. This estimate, say $\widehat{K}$, is used by the Isomap algorithm to produce an embedding into $\widehat{K}$-dimensional space. The residual variance produced by this operation is recorded. We then increment M by 1 and repeat the entire process. The algorithm terminates when the residual variance obtained is smaller than some tolerance parameter δ. A full length description is provided in Algorithm 6.

The essence of ML-RP is as follows. A sufficient number M of random projections is determined by a nonlinear procedure (i.e., sequential computation of Isomap residual variance) so that conventional manifold learning does almost as well on the projected dataset as the original. On the other hand, the random linear projections provide a faithful representation of the data in the geodesic sense. In this manner, ML-RP helps determine the number of rows that $\mathbf{\Phi}$ requires in order to act as an operator that preserves metric structure. Therefore, ML-RP can be viewed as an adaptive method for linear reduction of data dimensionality. It is only weakly adaptive in the sense that only the stopping criterion for ML-RP is determined by monitoring the nature of the projected data.

The results derived in Section 5.3 can be viewed as convergence proofs for ML-RP. The existence of a certain minimum number of measurements for any chosen error value δ ensures that eventually, M in the ML-RP algorithm is going to become high enough to ensure "good" Isomap performance. Also, due to the built-in parsimonious nature of ML-RP, we are ensured to not "overmeasure" the manifold, i.e., just the requisite numbers of projections of points are obtained.

5.5 Random Projections for Data Fusion

5.5.1 Joint manifolds

The theory and methods developed in Section 5.3 are applicable to high-dimensional data modeled by a single manifold $\mathcal{M} \in \mathbb{R}^N$. However, in scenarios involving sensor networks, multiple observations of the same event are often acquired simultaneously, resulting in the acquisition of interdependent signals that share a common parameterization. For example, a camera network might observe a single event from a variety of vantage points, where the underlying event is described by a set of common global parameters (such as the location and orientation of an object of interest). Similarly, when sensing a single phenomenon using multiple modalities, such as video and audio, the underlying phenomenon may again be described by a single parameterization that spans all modalities (such as when analyzing a video and audio recording of a person speaking, where both are parameterized by the phonemes being spoken). In both examples, all of the acquired signals are functions of the same set of parameters, i.e., we can write each signal as $f_j(\boldsymbol{\theta})$ where $\boldsymbol{\theta} \in \Theta$ is the same for all j.

Our contention is that we can obtain a simple model that captures the correlation between the sensor observations by matching the parameter values for the different manifolds observed by the sensors. More precisely, by simply concatenating points that are indexed by the same parameter value $\boldsymbol{\theta}$ from the different component manifolds, i.e., by forming $\mathbf{f}(\boldsymbol{\theta}) = [f_1(\boldsymbol{\theta}), f_2(\boldsymbol{\theta}), \ldots, f_J(\boldsymbol{\theta})]$, we obtain a new "parent" manifold that encompasses all of the component manifolds and shares the same parameterization. This structure captures the interdependencies between the signals in a straightforward manner. We can then apply the same manifold-based processing techniques that have been proposed for individual manifolds to the entire ensemble of component manifolds.

We formalize this intuition by defining the notion of a *joint manifold*. In order to simplify our notation, we will let $\boldsymbol{\mathcal{M}} = \mathcal{M}_1 \times \mathcal{M}_2 \times \cdots \times \mathcal{M}_J$ denote the *product manifold*. Furthermore, we will use the notation $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_J)$ to denote a J-tuple of points, or concatenation of J points, which lies in the Cartesian product of J sets (e.g., $\boldsymbol{\mathcal{M}}$).

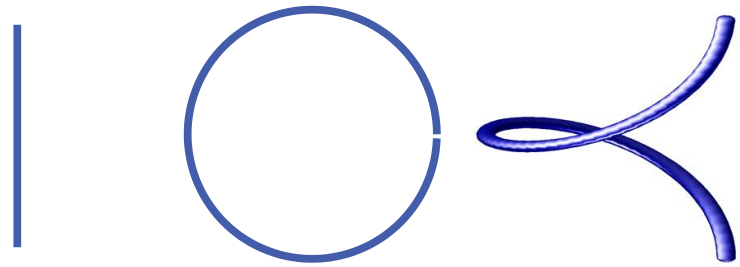
Definition 7. *Let $\{\mathcal{M}_j\}_{j=1}^J$ be an ensemble of J topological manifolds of equal dimension K. Suppose that the manifolds are homeomorphic to each other, in which case there exists a homeomorphism ψ_j between $\mathcal{M}_1$ and $\mathcal{M}_j$ for each j. For a particular set of $\{\psi_j\}_{j=2}^J$, we define the* **joint manifold** *as*

$$\mathcal{M}^* = \{\mathbf{x} \in \boldsymbol{\mathcal{M}} : \mathbf{x}_j = \psi_j(\mathbf{x}_1), 2 \leq j \leq J\}.$$

Furthermore, we say that $\{\mathcal{M}_j\}_{j=1}^J$ are the corresponding **component manifolds**.

Note that $\mathcal{M}_1$ serves as a common parameter space for all the component manifolds. Since the component manifolds are homeomorphic, this choice is ultimately arbitrary. In practice it may be more natural to think of each component manifold as being homeomorphic to some fixed K-dimensional parameter space Θ.

As a simple illustration, consider the one-dimensional manifolds in Fig. 5.2. Figures 5.2(a) and (b) show two isomorphic manifolds, where $\mathcal{M}_1 = (0, 2\pi)$ is an open interval, and $\mathcal{M}_2 = \{\psi_2(\theta) : \theta \in \mathcal{M}_1\}$ where $\psi_2(\theta) = (\cos(\theta), \sin(\theta))$, i.e., $\mathcal{M}_2 = S^1 \backslash (1, 0)$ is a circle with one point removed (so that it remains isomorphic to a line segment). In this case the joint manifold $\mathcal{M}^* =$



(a) $\mathcal{M}_1 \subseteq \mathbb{R}$: line segment

(b) $\mathcal{M}_2 \subseteq \mathbb{R}^2$: circle segment

(c) $\mathcal{M}^* \subseteq \mathbb{R}^3$: helix segment

Figure 5.2: sl A pair of isomorphic manifolds $\mathcal{M}_1$ and $\mathcal{M}_2$, and the resulting joint manifold $\mathcal{M}^*$.

$\{(\theta, \cos(\theta), \sin(\theta)) : \theta \in (0, 2\pi)\}$, illustrated in Fig. 5.2(c), is a helix. Notice that there exist other possible homeomorphic mappings from $\mathcal{M}_1$ to $\mathcal{M}_2$, and that the precise structure of the joint manifold as a submanifold of $\mathbb{R}^3$ is heavily dependent on the choice of this mapping.

5.5.2 Data fusion on joint manifolds

We leverage the joint manifold model (Definition 7) to develop a simple, efficient data fusion scheme. Given a network of J sensors, let $\mathbf{x}_j \in \mathbb{R}^N$, denote the signal acquired by sensor j, which is assumed to belong in a manifold $\mathcal{M}_j$, and let $\mathbf{x}$ denote the corresponding point in the joint manifold $\mathcal{M}^*$. Rather than forming the vector $\mathbf{x}$, one could potentially estimate a K-dimensional parameter vector $\widehat{\boldsymbol{\theta}}_j$ via the nonlinear mapping of $\mathbf{x}_j$ corresponding to the manifold $\mathcal{M}_j$. By transmitting the $\widehat{\boldsymbol{\theta}}_j$ at a central location, we would obtain a data representation of dimension JK. For example, in [145] this technique is implemented using a Laplace-Beltrami embedding to extract the parameter values for each signal.

However, the Laplace-Beltrami approach can be sometimes problematic. By simply concatenating each $\widehat{\boldsymbol{\theta}}_j$, this approach essentially ignores the joint manifold structure present in the data, which is evident due to the fact that in an ideal setting the same K parameters will be obtained from each of the J sensors. Moreover, given noisy estimates for $\widehat{\boldsymbol{\theta}}_j$, it is not obvious how to most effectively integrate the $\widehat{\boldsymbol{\theta}}_j$ to obtain a single joint K-dimensional representation. Finally, while this approach eliminates the dependence on N, the cost of transmission of the parameters $\widehat{\boldsymbol{\theta}}_j$ still suffers from a linear dependence on J.

In order to address this problem, we exploit the joint manifold structure to develop a more efficient fusion scheme. Our approach is based on the fundamental principles described in Section 5.2. We will aim to compute a dimensionally reduced representation of $\mathbf{x}$ denoted $\mathbf{y} = \boldsymbol{\Phi}\mathbf{x}$, where $\boldsymbol{\Phi}$ is a standard linear projection operator. Since the operator is linear, we can take *local* projections of the images acquired by each sensor, and still calculate the *global* projections of $\mathbf{x}$ in a distributed fashion. Let each sensor calculate $\mathbf{y}_j = \boldsymbol{\Phi}_j \mathbf{x}_j$, with the matrices $\boldsymbol{\Phi}_j \in \mathbb{R}^{M \times N}, 1 \leq j \leq J$. Then, by defining the $M \times JN$ matrix $\boldsymbol{\Phi} = [\boldsymbol{\Phi}_1 \ \boldsymbol{\Phi}_2 \ \cdots \ \boldsymbol{\Phi}_J]$, the global projections $\mathbf{y} = \boldsymbol{\Phi}\mathbf{x}$ can be obtained

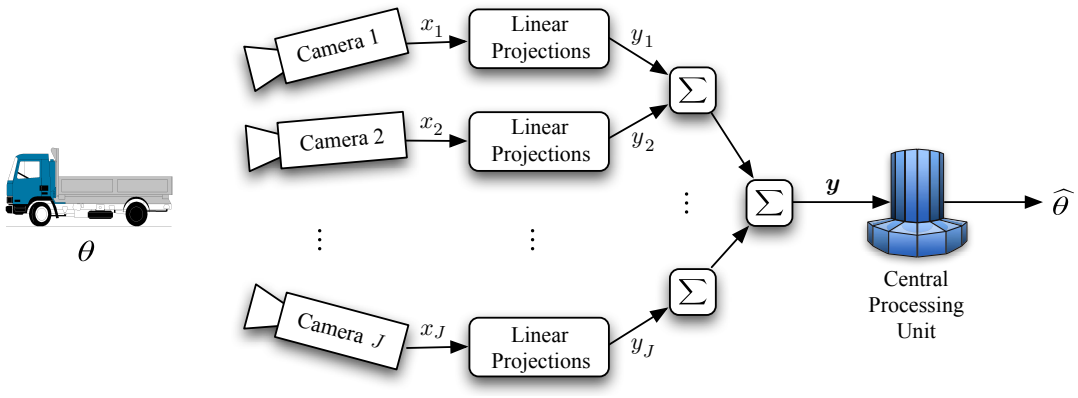


Figure 5.3: *Distributed data fusion using linear projections in a camera network.*

by

$$\begin{aligned}\mathbf{y} = \mathbf{\Phi}\mathbf{x} &= [\mathbf{\Phi}_1 \quad \mathbf{\Phi}_2 \quad \cdots \quad \mathbf{\Phi}_J][\mathbf{x}_1^T \quad \mathbf{x}_2^T \quad \cdots \quad \mathbf{x}_J^T]^T \\ &= \mathbf{\Phi}_1\mathbf{x}_1 + \mathbf{\Phi}_2\mathbf{x}_2 + \cdots + \mathbf{\Phi}_J\mathbf{x}_J.\end{aligned}$$

Thus, the final measurement vector can be obtained by simply *adding independent projections* of the signals acquired by the individual sensors. This gives rise to the *compressive data fusion* protocol illustrated in Fig. 5.3. Suppose the individual sensors are associated with the nodes of a binary tree of size J, where the edges represent communication links between nodes. Let the root of the tree denote the final destination of the fused data (the central processing unit). Then the fusion process can be represented by the flow of data from the leaves to the root, with a binary addition occurring at every parent node.

The main challenge in designing such a scheme is the choice of a suitable matrix $\mathbf{\Phi}$. Fortunately in this case, we can leverage the random projections approach developed in Section 5.2. In particular, Theorem 4 asserts that data belonging to a smooth manifold in $\mathbb{R}^N$ can be stably embedded in a subspace that is only logarithmic in N. In the multiple-manifold context, this implies that if the joint manifold has bounded condition number, then we can project the joint data into a random subspace of dimension that is only logarithmic in J and N and still approximately preserve the manifold structure. This is formalized in the following Corollary, which is a direct consequence of Theorem 4.

Corollary 3. *Let $\mathcal{M}^*$ be a compact, smooth, Riemannian joint manifold in a JN-dimensional space with condition number $1/\tau^*$. Let $\mathbf{\Phi}$ denote an orthogonal linear mapping from $\mathcal{M}^*$ into a random M-dimensional subspace of $\mathbb{R}^{JN}$. Let $M = O(K\log(JN/\tau^*)/\delta^2)$. Then, with high probability, the geodesic and Euclidean distances between any pair of points on $\mathcal{M}^*$ are preserved up to distortion δ under $\mathbf{\Phi}$.*

In other words, we can obtain a faithful embedding of the joint manifold via a representation of dimension only $O(K\log JN)$. This represents a massive improvement over the original JN-dimensional representation of the aggregate data.

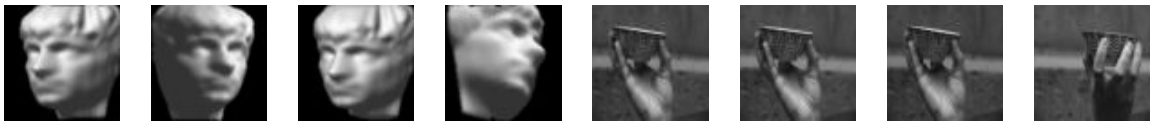


Figure 5.4: *Standard databases. Ambient dimension for the face database $N = 4096$; ambient dimension for the hand rotation databases $N = 3840$.*

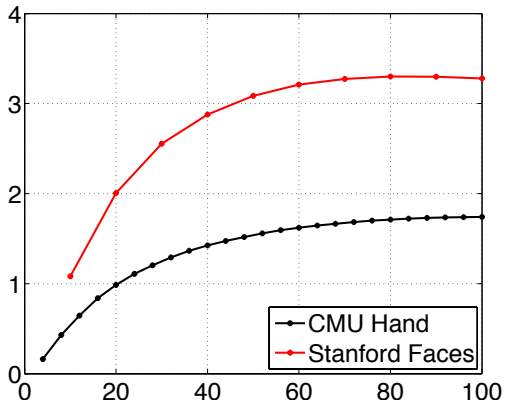


Figure 5.5: *Performance of ID estimation, using GP as a function of random projections, for the datasets in Fig. 5.4. The estimates of the intrinsic dimensions converge to the 'true' estimates with merely $M = 100$ random linear measurements.*

Further, this low-dimensional representation can be efficiently calculated in-network using the simple tree-structured data routing scheme. The total communication bandwidth required for our compressive data fusion scheme is $O(M \log J)$ .Therefore, we obtain that when using random projections, the dependence of the required bandwidth on J is merely $O(\log^2 J)$. This offers a significant improvement from previous data fusion methods that necessarily require the communication bandwidth to scale linearly with the number of cameras. Our proposed linear fusion approach via linear projections integrates the network transmission and data fusion steps in a fashion similar to the protocols discussed in randomized gossiping [146] and compressive wireless sensing [147].

5.6 Experiments

We now test the performance of our proposed random projection-based approaches for efficient learning of high-dimensional data, and demonstrate the considerable benefits of our approach. First, we focus our attention on the problem of intrinsic dimension (ID) estimation. We consider two common datasets found in the literature on dimension estimation — the face database [60], and the hand rotation database [148]. Example images of both datasets are shown in Fig. 5.4. The face database is a collection of 698 artificial snapshots of a face ($N = 64 \times 64 = 4096$) varying under 3 degrees of freedom: 2 angles for pose and 1 for lighting dimension. The signals are therefore believed to reside on a 3D manifold in an ambient space of dimension 4096. The hand rotation

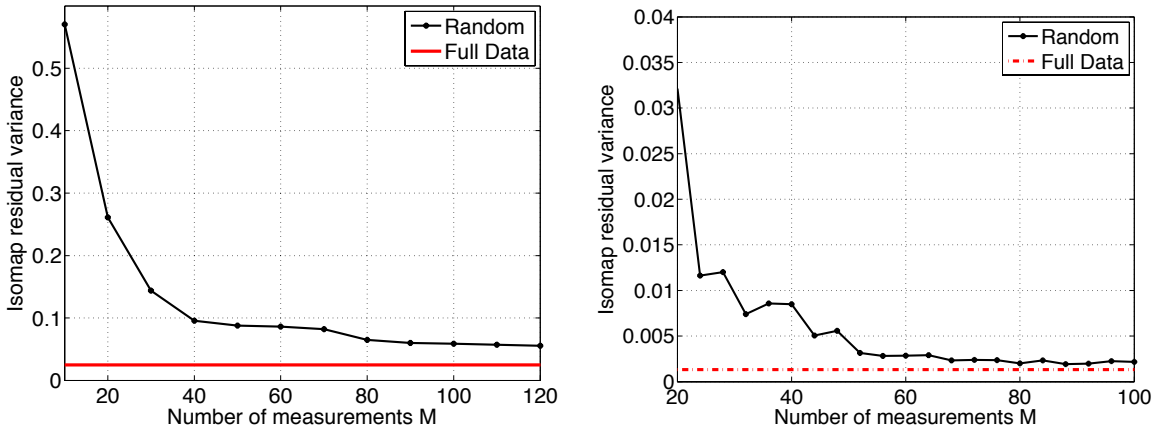


Figure 5.6: *Performance of ML-RP for the datasets in Fig. 5.4. (left) ML-RP on the face database ($N = 4096$). Good approximates are obtained for $M > 50$. (right) ML-RP on the hand rotation database ($N = 3840$). For $M > 60$, the Isomap variance is indistinguishable from the variance obtained in the ambient space.*

database is a set of 90 images ($N = 64 \times 60 = 3840$) of rotations of a hand holding an object. Although the image appearance manifold is ostensibly one-dimensional, estimators in the literature always overestimate its ID [143].

We numerically test the performance of the GP algorithm for intrinsic dimension (ID) estimation. Random projections of each sample in the databases were obtained by computing the inner product of the image samples with an increasing number of rows M of the random orthoprojector $\mathbf{\Phi}$. The results of this approach are displayed in Fig. 5.5. We observe that the ID estimate stabilizes quickly with increasing number of projections. This is in accordance with Theorem 5, which implies that with increasing M, the ID estimate using the random projections closely approximates the estimate using the original high-dimensional data.

Figure 5.6 indicates the performance of our proposed ML-RP algorithm for these datasets. We note that in the case of the face database, for $M > 60$, the Isomap variance on the randomly projected points closely approximates the variance obtained with full image data. This behavior of convergence of the variance to the best possible value is even more sharply observed in the hand rotation database, in which the two variance curves are indistinguishable for $M > 60$. These results are particularly encouraging and demonstrate the validity of the claims made in Section 5.3.

Next, we provide a variety of results using both simulated and real data that demonstrate the significant gains obtained by using the joint manifold model, both with and without the use of random projections. All manifold learning results have been generated using Isomap (Algorithm 2). For ease of presentation, all of our experiments are performed on two-dimensional (2D) translation manifolds ($K = 2$) that are isomorphic to a closed rectangular subset of $\mathbb{R}^2$. Thus, ideally the 2D embedding of the data should resemble a rectangular grid of points that correspond to the samples of the joint manifold in high dimensional space.

We test our proposed data fusion scheme (Fig. 5.3) using image datasets from a real-world

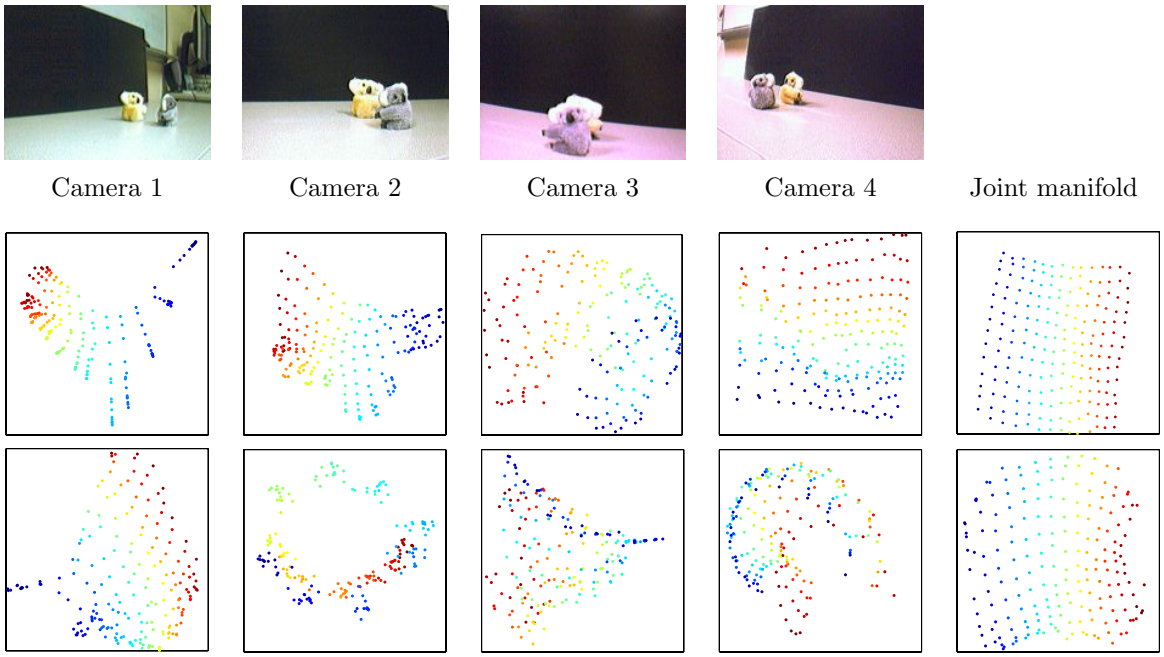


Figure 5.7: *(top) Sample images of 2 koalas moving along individual 1-D paths, yielding a 2D manifold; (middle) 2D embeddings of the dataset learned via Isomap from $N = 76800$ pixel images; (bottom) 2D embeddings of the dataset learned from $M = 2400$ random projections. Learning the joint manifold yields a much improved 2D embedding.*

camera network. The images are obtained from a network of four Unibrain Fire-i$^{\text{TM}}$ OEM Firewire board cameras. Each camera has resolution $N = 320 \times 240 = 76800$ pixels. The data comprises $J = 4$ different views of the independent motions of 2 toy koalas along individual 1-D paths, yielding a 2D combined parameter space. This data suffers from real-world artifacts such as fluctuations in illumination conditions and variations in the pose of the koalas; further, the koalas occlude one another in certain views or are absent from certain views depending on the particular vantage point.

Sample images and 2D embedding results are displayed in Fig. 5.7. We observe that the best embedding is obtained by using the modified version of Isomap for learning the joint manifold. To test the effectiveness of the data fusion method described in Section 5.5, we compute $M = 2400$ random projections of each image and sum them to obtain a randomly projected version of the joint data and repeat the above experiment. The dimensionality of the projected data is only 3% of the original data; yet, we see very little degradation in performance, thus displaying the effectiveness of random projection-based fusion.

We consider a situation where we are given a set of training data consisting of images of a target moving through a region along with a set of test images of the target moving along a particular trajectory. Note that this is an "unsupervised" learning task in the sense that we do not explicitly incorporate any known information regarding the locations of the cameras or the parameter space describing the target's motion; this information must be implicitly learned from the training data. The training images comprise $J = 4$ views of a coffee mug placed at different positions on an

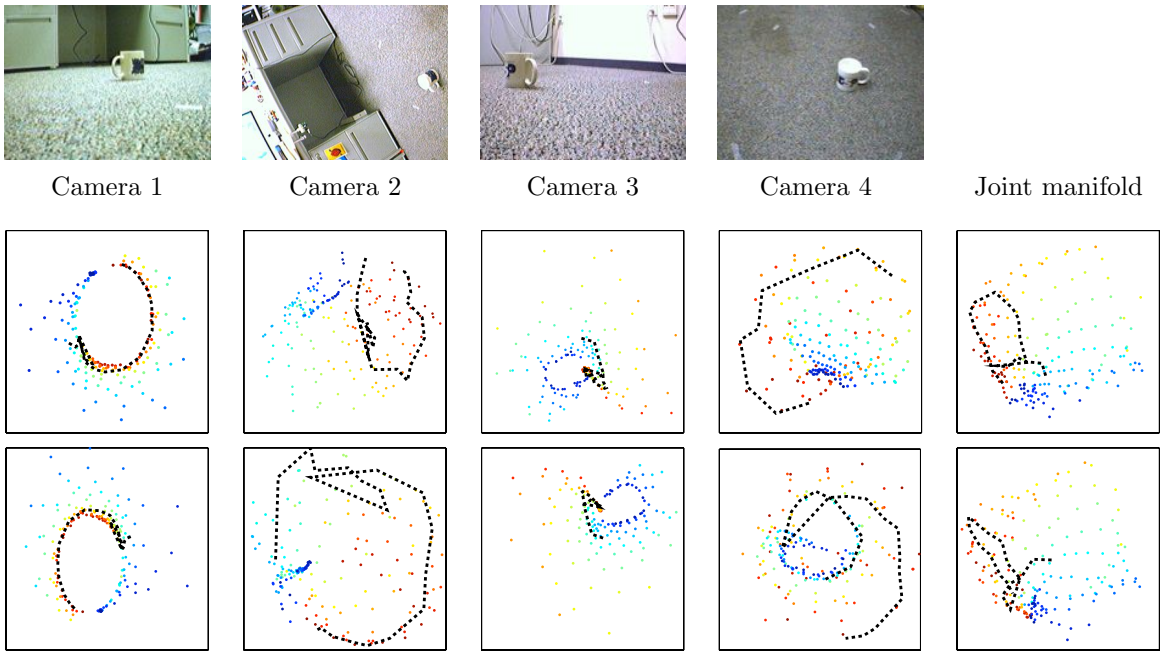


Figure 5.8: *(top) Sample images of the 2D movement of a coffee mug; (middle) 2D embeddings of the dataset learned via Isomap from $N = 76800$ pixel images; (bottom) 2D embeddings of the dataset learned via Isomap from $M = 4800$ random projections. The black dotted line corresponds to an "R"-shaped trajectory in physical space.*

irregular rectangular grid. Example images from each camera are shown in Fig. 5.8. For the test data, we translate the coffee mug so that its 2D path traces out the shape of the letter "R". We aim to recover this shape using both the test and training data. To solve this problem, we attempt to learn the 2D Isomap embedding of the image manifolds corresponding to each camera, as well as the joint manifold formed by aggregating data from all cameras. The learned embeddings are shown in Fig. 5.8. As is visually evident, learning the data using any one camera yields very poor results; however learning the joint manifold helps discern the 2D structure to a much better degree. In particular, the "R" trajectory in the test data is correctly recovered only by learning the joint manifold.

Finally, we repeat the above procedure using $M = 4800$ random projections of each image, and fuse the data by summing the measurement vectors. While the recovered trajectory of the anomalous (test) data suffers some degradation in visual quality, we observe comparable 2D embedding results for the individual and joint manifolds as with the original data set. Since the dimensionality of the projected data is merely 6% that of the original data set, this would translate to significant savings in communication costs in a real-world camera network.

5.7 Discussion

The method of random projections is thus a powerful tool for enabling efficient *inference* algorithms for high-dimensional data that are well-modeled by nonlinear manifolds. The motivation for developing theory and algorithms that involve random measurements of high-dimensional data is significant, particularly due to the increasing attention that Compressive Sensing (CS) has received. It is now possible to think of settings involving a huge number of low-power devices that inexpensively capture, store, and transmit a very small number of measurements of high-dimensional data. Our proposed algorithm, ML-RP (Algorithm 6) is applicable in all such settings. In situations where the bottleneck lies in the transmission of the data to the central processing node, ML-RP provides a simple solution to the manifold learning problem and ensures that with a minimum amount of available information, effective manifold learning can be performed.

In order to handle multi-sensor scenarios where multiple signal ensembles are governed by a small number of common parameters, we propose a new geometric model called the *joint manifold*. We leverage this model, coupled with our theory and algorithms using random projections, to form a simple data fusion scheme. Our proposed framework for data fusion simply employs independent random projections of each sensor that are then accumulated to obtain an accurate low-dimensional representation of the joint manifold. Joint manifold fusion via random projections is *universal* in that the projections do not dependent on the specific structure of the manifold. Thus, our sensing techniques need not be replaced for these extensions; only our underlying models (hypotheses) are updated.

CHAPTER 6

Learning Manifolds in the Wild

6.1 Setup

In this chapter, we consider the problem of *robust* manifold modeling and processing of real-world image ensembles. Manifold-based models have long been used for applications involving data ensembles that can be described by only a few degrees of freedom. The promise of such models lies in their ability to potentially break the so-called "curse of dimensionality", a common problem in most practical machine learning tasks. Consequently, the last decade has witnessed great theoretical and algorithmic advances in this regard, and manifold models have been successfully applied to tasks such as data visualization, parameter estimation, transductive learning, and compact data representations [60, 61, 71, 72].

However, the significant theoretical advances in manifold-based methods have not led to commensurate success in practice, particularly in the context of modeling and processing large-scale image data. The reasons for this stem from two fundamental challenges:

1. **Lack of isometry:** A common desideratum for several manifold-based algorithms is that the underlying manifold is isometric to the underlying parameter space, i.e., small changes in the articulation parameter $\boldsymbol{\theta}$ generate images that are "nearby" in terms of Euclidean distance. Unfortunately, this assumption breaks down for anything except the simplest of IAMs. However, extensive calculations [64] have shown that for anything more complicated than a simple white object moving over a black background, local isometry does not hold.

2. **Nuisance variables:** In addition to the small number of degrees of freedom in the articulations of interest, real-world images ensembles often exhibit a potentially large number

This work is in collaboration with Richard G. Baraniuk and Aswin C. Sankaranarayanan [149].

Figure 6.1: *An image ensemble gathered from the wild. Example images of the Notre Dame Cathedral gathered from FlickR. Such a real-world image ensemble cannot be easily modeled via a strictly low-dimensional parametric representation; occlusions are significant, illlumination variations are dramatic, and imaging artifacts such as varying field-of-view, skew, and white balance abound. As a consequence, conventional manifold learning methods fail when applied to such ensembles.*

of other, nuisance articulations, such as illumination variations, changing backgrounds and clutter, and occlusions due to foreground objects. See Fig. 6.1 for an illustrative example.

This mismatch between theoretical assumptions and practical realities has diminished the impact of manifold models for real-world machine learning and vision problems.

In this chapter, we develop a new framework for manifold-based modeling, learning, and processing of large-scale image ensembles that directly addresses the two challenges. Our contribution is threefold. First, to address the isometry challenge, we rigorously prove that the classical *Earth Mover's Distance* (EMD) between images can be used to establish isometry for image ensembles generated by translations and rotations of a reference image. This result makes no restrictive assumptions and holds even when the images under consideration are highly textured grayscale images. To the best of our knowledge, this is the first analytical result proving the isometry of manifolds of generic grayscale images.

Second, to address the nuisance variable challenge, we advocate a new image representation for manifold modeling, learning, and processing. Given a set of articulating images, we represent each image using a set of *local features* (or keypoints). Such an approach is ubiquitous in practical computer vision approaches. A keypoint typically consists of a 2D *location* in the image domain and a higher-dimensional *descriptor* summarizing the local statistics of the grayscale values of the image. We will require that the keypoint locations and descriptors satisfy certain stability criteria (explained further in Section 6.4). Our running example will be the image features generated by the well-known Scale Invariant Feature Transform (SIFT) [150], but other image features are also

possible within this framework. Under this new representation, we show that the transformed set of images can be viewed as a low-dimensional manifold that we dub the *keypoint articulation manifold* (KAM). In fact we prove that, under a suitable modification of the EMD metric, the KAM is smooth and isometric to the underlying parameter space. By moving to this alternate representation, we implicitly promote robustness to various nuisance parameters (such as varying illumination, backgrounds, occlusions, and clutter). Therefore, our proposed KAM modeling approach alleviates both of the challenges encountered in practical applications.

Third, to mitigate computational complexity concerns related to the EMD, we propose a fast EMD approximation based on similarity kernels between the keypoint representations. We validate the approximation on several real datasets and manifold-based learning problems and demonstrate improved manifold embeddings, improved parameter estimation on affine articulation manifolds using gradient descent, and a fast, efficient, and automatic organization of large unordered collections of photographs.

The rest of this chapter is organized as follows. In Section 6.2, we review the existing literature on the nonlinear dimensionality reduction of image manifolds. In particular, we highlight some efforts geared towards addressing some of the fundamental challenges towards practical use of image manifolds, and discuss their limitations. In Section 6.3, we describe how the EMD ensures isometry of manifolds for simple classes of articulations. In Section 6.4, we extend the EMD to be applicable to a local feature-based image representation that enables robustness to undesirable articulations. In Section 6.5, we illustrate the performance of our approach on a range of manifold modeling and processing applications, and validate our technique on a number of image ensembles. In Section 6.6, we conclude with a discussion and highlight some directions for future research.

6.2 Related Work

We adopt two complementary representations for images. First, we can model images as *continuous* functions on $\mathbb{R}^2$, i.e., $I : \mathbb{R}^2 \mapsto \mathbb{R}$. Second, we can model images as *discretized* functions defined over a domain of size $n \times n$. In such situations, the ensemble of images are modeled as points in $\mathbb{R}^N$, where $N = n^2$. We will use these two representations interchangeably when the context is clear.

We are interested in image ensembles that are generated by varying an articulation parameter $\boldsymbol{\theta} \in \Theta$. If Θ is a space of dimension K, then the ensemble of images forms a K-dimensional nonlinear image articulation manifold (IAM) $\mathcal{M} = \{I_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\} \subset \mathbb{R}^N$. As detailed in Section 2.5, extensive calculations have shown that IAMs formed by images with sharp edges are, in general, *non-differentiable*, and *non-isometric* to the underlying parameter space [64]. This intrinsic nature of IAMs can severely degrade the performance of manifold learning methods when applied to real-world large scale image ensembles.

6.2.1 Transport operators

An attempt to resolve both the lack of smoothness and isometry in image manifolds is via the machinery of *transport operators*. Transport operators are functions $f_{\boldsymbol{\theta}} : \mathbb{R}^2 \mapsto \mathbb{R}^2$ that enable the *motion* along the surface of an IAM from one image I_0 to another $I_{\boldsymbol{\theta}}$, as specified by the algebraic

equation:

$$I_{\boldsymbol{\theta}}(\boldsymbol{x}) = I_0(\boldsymbol{x} + f_{\boldsymbol{\theta}}(\boldsymbol{x})). \tag{6.1}$$

In essence, transport operators serves as a *generative model* for images on the IAM.

For many IAMs, a modeling approach based on transport operators can be elegant and analytically tractable. For example, in the case where the IAM $\mathcal{M}$ is generated by affine articulations of a fixed template image I_0, the set of corresponding transport operators forms a *Lie group*.Based on this intuition, [151] develop a manifold learning framework using a *matrix exponential*-based generative model and demonstrate improved manifold-modeling performance. In addition to affine articulations, other common examples of analytically tractable transport operators are the projective transformations and the set of diffeomorphisms [152–154].

Many IAMs of practical interest, including the manifolds corresponding to 3D pose articulations and non-rigid deformations, possess no analytical algebraic structure. In previous work, we have extended the concept of transport operators for non-algebraic articulations such as 3D pose articulations [155]. The key observation behind our approach is that the set of transport operators *itself* forms a low-dimensional smooth manifold. Further, in certain situations, the manifold of transport operators can be easily computed using the optical flow between image pairs [156]. Such an approach enables a simple representation for IAMs that is suitable for applications such as image synthesis [155].

While approaches based on transport operators can help solve the lack of isometry of IAMs to some extent, a key drawback is their inability to handle nuisance articulations such as illuminations, background changes, and occlusions. In the presence of such nuisance articulations, the fundamental model defined in (6.1) is violated. As a consequence, their performance on images ensembles obtained in the wild is often poor.

6.2.2 Local image features

Modern computer vision algorithms often eschew the pixel intensity representation for a more convenient, *feature-based* representation. Such a feature-based image modeling approach has found widespread use in a multitude of practical applications, including object recognition [157], multi-view 3D scene reconstruction [158], and manipulating and visualizing massive photo collections [159]. For an introduction to image features and their properties, see [160] and [161].

Perhaps the most popular feature-based representation of images is obtained by the *Scale Invariant Feature Transform* (SIFT) [150]. The core idea underlying the SIFT technique is the notion of *scale space* [162]. The scale space of an image I is the 3D scalar-valued function $L : \mathbb{R}^2 \times \mathbb{R} \mapsto \mathbb{R}$ obtained by convolving I with an isotropic Gaussian smoothing kernel of scale s so that

$$L(\boldsymbol{x}, s) = \phi_s * I. \tag{6.2}$$

Rich information about an image can be gleaned by analyzing the Laplacian of the scale space of the image, or $\nabla^2 L(\boldsymbol{x}, s)$. Indeed, extensive testing [160] has shown that the locations of maxima and minima of $\nabla^2 L(\boldsymbol{x}, s)$ (denoted by a list of 2D locations and scales $S = \{\boldsymbol{x}^i, s^i\}$) are extremely stable to small affine deformations of the image. The SIFT technique leverages this property of scale space to extract distinctive features from images.

Numerically, the SIFT technique proceeds as follows. An image I is operated upon to obtain a set of 2D locations called *keypoint locations* $\{\boldsymbol{x}^i, i = 1, \ldots, M\}$; these are precisely the extrema of the Laplacian of the scale space of I. Each keypoint location $\boldsymbol{x}^i$ is assigned a scale s^i, and an orientation $\boldsymbol{\theta}^i$. Once the set of keypoint locations are identified, certain special image statistics around each keypoint are computed and aggregated in the form of histograms. Such histograms are stored as high-dimensional vectors known as *keypoint descriptors* $\{\boldsymbol{f}^i, i = 1, \ldots, M\}$.

It has been both theoretically and empirically demonstrated that the SIFT keypoint locations are *covariant* to affine articulations, while the SIFT keypoint descriptors are *invariant* to a wide range of imaging parameters, including translations, in-plane rotations, scale, and illumination changes [150, 163]. Let I_A and I_B be two images with keypoints given by $S(I_A) = \{\boldsymbol{x}_A^i\}$ and $S(I_B) = \{\boldsymbol{x}_B^j\}$, respectively. If the two images are related by an affine transformation (Z, t), then the keypoints are related by the same affine transformation (ignoring quantization and boundary artifacts):

$$I_B(\boldsymbol{x}) = I_A(Z\boldsymbol{x} + \boldsymbol{t}) \implies \forall i, \exists j \text{ such that } \boldsymbol{x}_B^j = Z\boldsymbol{x}_A^i + \boldsymbol{t}. \tag{6.3}$$

Therefore, by obtaining one-to-one correspondences between the keypoint descriptors of I_A and I_B, we can solve for the affine transformation (Z, t) linking the two images.

We have nominally chosen to focus on the SIFT as our flagship approach for generating image features, but other feature extraction techniques can also be applied in the framework developed below (for example, see [164–166]). In general, we will require that any such technique should yield image feature keypoints whose locations are covariant to the articulations of interest, and whose descriptors are invariant to the keypoint location, as well as other nuisance articulations.[1] The covariance-invariance properties help mitigate several phenomena such as unknown illuminations, occlusion, and clutter as detailed further below in Sections 6.4 and 6.5.

The large majority of manifold learning methods do not leverage the feature-based approach for representing images. To the best of our knowledge, the only reported manifold learning method that explicitly advocates feature-based image representations is the *Local Features* approach [167]. Given a collection of images, this approach extracts a set of local features from each of the images, and then learns a low-dimensional parametric embedding of *each* extracted feature. This embedding is constrained to preserve the spatial configuration of features. Further, similarity kernels are used to construct similarities on the keypoint locations and descriptors, and embeddings of the keypoints are learnt. This method has been shown to be robust to illumination, occlusions, and other artifacts, and thus shares many of the goals of our proposed approach. However, its theoretical development is somewhat ad hoc, its computational costs are potentially high, and the reported applications are mainly restricted to object detection and classification. We will discuss and compare our results to the Local Features approach in detail in Section 6.5.

[1] Naturally, the trivial (zero) feature descriptor also satisfies this invariance requirement. Our theoretical results below will continue to be valid for such degenerate cases; however, a meaningful feature descriptor that concisely represents local image statistics is obviously the better choice in practice.

6.3 Manifold Isometry via the Earth Mover's Distance

The central results of [64] advocating the multiscale smoothing approach for enabling manifold isometry were derived based on the assumption that images are modeled as functions defined on $\mathbb{R}^2$ equipped with the L_2-norm. However, this modeling assumption is comes up short in a key respect: L_2-distances between images are known to be poorly correlated with perceptual differences between images. For example, given images of a single translating white dot on a black background, the L_2-distance between any pair of images remains constant regardless of the translation parameters of the images.

6.3.1 The Earth Mover's Distance (EMD)

To address the pitfall caused by L_2-distances, researchers have proposed a multitude of alternate, perceptually meaningful distance measures on images. An important and useful metric used in image retrieval and analysis is the Earth Mover's Distance (EMD) [168]. Classically, the EMD is defined between distributions of mass over a domain, and represents the minimal amount of *work* needed to transform one distribution into another. In this context, the amount of work required to move a unit of mass from a point $\boldsymbol{x}_1 \in \mathbb{R}^2$ to a point $\boldsymbol{x}_2 \in \mathbb{R}^2$ is equal to the L_2-norm between $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$.

For ease of exposition we will assume that images are defined over a discrete grid in $\mathbb{R}^2$, while noting that the results hold *mutatis mutandis* for continuous domain images. Formally, consider images I_1, I_2 as non-negative functions defined on a domain of size $n \times n$. Define a *feasible flow* as a function $\gamma : [n]^2 \times [n]^2 \to \mathbb{R}_+$ that satisfies the mass conservation constraints, i.e., for any pair of pixel locations $\boldsymbol{x}_i, \boldsymbol{y}_j \in [n]^2$,

$$\sum_{\boldsymbol{y}_k \in [n]^2} \gamma(\boldsymbol{x}_i, \boldsymbol{y}_k) = I_1(\boldsymbol{x}_i), \qquad \sum_{\boldsymbol{x}_k \in [n]^2} \gamma(\boldsymbol{x}_k, \boldsymbol{y}_j) = I_2(\boldsymbol{y}_j).$$

Then, we define

$$EMD(I_1, I_2) = \min_{\gamma} \sum_{\boldsymbol{x}_i, \boldsymbol{y}_j \in [n]^2} \gamma(\boldsymbol{x}_i, \boldsymbol{y}_j) \|\boldsymbol{x}_i - \boldsymbol{y}_j\|_2, \tag{6.4}$$

as the minimum cost flow from X to Y over all feasible flows. If the sum of the absolute values of the intensities of X and Y are equal, i.e., if $\|X\|_1 = \|Y\|_1$, then it can be shown that $EMD(X, Y)$ is a valid metric on the space of images. In this section, we will assume the equality of the ℓ_1 norms of X and Y; however, the metric property of the EMD holds even when this assumption is relaxed [168]. Unless otherwise specified we will assume that the EMD is always computed between images of equal ℓ_1 norm.

The EMD provides a powerful new angle for studying the geometric structure of image manifolds. As opposed to modeling images as functions in $L_2(\mathbb{R}^2)$, we instead represent images as elements of the normed space $L_{EMD}(\mathbb{R}^2)$. Under this geometry, we can prove the isometry of a much larger class of image ensembles; we discuss now some representative examples.

6.3.2 Case study: Translation manifolds

We prove the *global* isometry of image manifolds in $L_{EMD}(\mathbb{R}^2)$ formed by arbitrary translations of a generic grayscale image. Consider a grayscale image I_0, and denote $\mathcal{M}_{\text{trans}}$ as the IAM generated by 2D translations of I_0, where $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^2$ represents the translation parameter vector:

$$\mathcal{M} = \{I : I(\boldsymbol{x}) = I_0(\boldsymbol{x} - \boldsymbol{\theta}),\ \boldsymbol{\theta} \in \Theta\}.$$

In order to avoid boundary and digitization effects, we will assume that the space of translation parameters Θ is compact, that the image has been sufficiently zero-padded, and that the images are of high resolution. It follows that the ℓ_1 norm of any image belonging to $\mathcal{M}_{\text{trans}}$ remains constant.

Proposition 2. *For an arbitrary base image I_0, the translation manifold $\mathcal{M}_{\text{trans}}$ is* globally isometric *to the parameter space Θ under the EMD metric.*

Proof: Consider any pair of images

$$I_1(\boldsymbol{x}) = I_0(\boldsymbol{x} - \boldsymbol{\theta}_1), \quad I_2(\boldsymbol{x}) = I_0(\boldsymbol{x} - \boldsymbol{\theta}_2)$$

that are elements of $\mathcal{M}_{\text{trans}}$. We will prove that $EMD(I_1, I_2)$ is proportional to the L_2 distances between the corresponding parameter vectors $\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_2$. Let $\check{\boldsymbol{x}}$ denote the *center of mass* of the image $I(\boldsymbol{x})$:

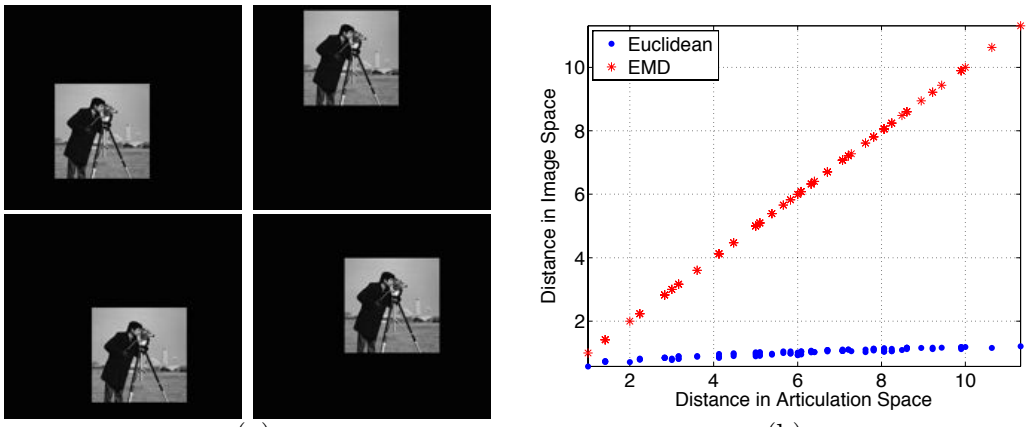
$$\check{\boldsymbol{x}} = \frac{1}{\|I\|_1} \sum_{\boldsymbol{x}_k \in [n]^2} \boldsymbol{x}_k I(\boldsymbol{x}_k).$$

Then, we have the following relations between the centers of mass of I_1, I_2 and *any* feasible flow f:

$$\begin{aligned}
\|\check{\boldsymbol{x}}_1 - \check{\boldsymbol{x}}_2\|_2 &= \left\| \frac{\sum_i \boldsymbol{x}_i I_1(\boldsymbol{x}_i)}{\|I_1\|} - \frac{\sum_j \boldsymbol{y}_j\ I_2(\boldsymbol{y}_j)}{\|I_2\|_1} \right\|_2 \\
&= C \left\| \sum_i \boldsymbol{x}_i I_1(\boldsymbol{x}_i) - \sum_j \boldsymbol{y}_j I_2(\boldsymbol{y}_j) \right\|_2 \\
&= C \left\| \sum_i \boldsymbol{x}_i \sum_k \gamma(\boldsymbol{x}_i, \boldsymbol{y}_k) - \sum_j \boldsymbol{y}_j \sum_k \gamma(\boldsymbol{x}_k, \boldsymbol{y}_j) \right\|_2 \\
&= C \left\| \sum_{i,k} \gamma(\boldsymbol{x}_i, \boldsymbol{y}_k)\boldsymbol{x}_i - \sum_{j,k} \gamma(\boldsymbol{x}_k, \boldsymbol{y}_j)\boldsymbol{y}_j \right\|_2 = C \left\| \sum_{i,j} \gamma(\boldsymbol{x}_i, \boldsymbol{y}_j)(\boldsymbol{x}_i - \boldsymbol{y}_j) \right\|_2 \\
&\leq C \sum_{i,j} \gamma(\boldsymbol{x}_i, \boldsymbol{y}_j) \left\| \boldsymbol{x}_i - \boldsymbol{y}_j \right\|_2,
\end{aligned}$$

where the last inequality is a consequence of the triangle inequality. Taking the infimum over all possible feasible flows, we have that

$$\|\check{\boldsymbol{x}}_1 - \check{\boldsymbol{x}}_2\|_2 \leq C \cdot EMD(I_1, I_2). \tag{6.5}$$



(a) (b)

Figure 6.2: *(a) Sample images from a translation manifold. (b) Variation of the Euclidean distance and the EMD as a function of the distance in the articulation space. The EMD correlates linearly with articulation distance for the entire range of articulations (global isometry).*

However, in the case of images that are 2D translations of one another, there always exists a feasible flow that *achieves* this infimum. This can be represented by the set of flows parallel to $\check{\boldsymbol{x}}_1 - \check{\boldsymbol{x}}_2$ originating from the pixel $\boldsymbol{x}_i$ and terminating at the corresponding $\boldsymbol{y}_j$. We simply rewrite the vector $\check{\boldsymbol{x}}_1 - \check{\boldsymbol{x}}_2$ as the difference in translation vectors $\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2$, thereby implying that

$$EMD(I_1, I_2) \propto \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_2.$$

Global isometry of $\mathcal{M}_{\text{trans}}$ is an immediate consequence. □

We numerically illustrate the validity of Proposition 2 in Fig. 6.2. Figure 6.2(a) displays several sample images from the manifold formed by translations of the well-known *Cameraman* test image. We form 100 example pairs of such images, record the distance between the translation parameter vectors (the "distance in articulation space"), and compute the Euclidean (ℓ_2) distance and EMD between the corresponding images. We compute the EMD using the FastEMD solver [169]. Figure 6.2(b) clearly indicates that the ℓ_2 distance is largely uninformative with respect to the articulation distance, while the EMD almost perfectly correlates with the articulation distance over the entire range of translations (global isometry).

6.3.3 Case study: Rotation manifolds

We prove the *local* isometry of image manifolds formed by rotations of a generic grayscale image. The IAM $\mathcal{M}_{\text{rot}}$ is generated by pivoting an image I_0 by an angle $\theta \in \Theta \subset [-\pi, \pi]$, around a fixed point in $\mathbb{R}^2$. We assume without loss of generality that the pivot point is the origin. Then, the manifold $\mathcal{M}_{\text{rot}}$ is given by

$$\begin{aligned} \mathcal{M} &= \{I : I(\boldsymbol{x}) = I_0(R_\theta \boldsymbol{x}),\ \theta \in \Theta\}, \quad \text{where} \\ R_\theta &= \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}, \end{aligned}$$

i.e., R_θ is a orthonormal rotation matrix. Once again, we assume that the images are sufficiently zero padded and that their ℓ_1 norms remain constant.

Proposition 3. *For an arbitrary base image I_0, the rotation manifold $\mathcal{M}_{\text{rot}}$ is* locally isometric *to the parameter space Θ under the EMD metric.*

Proof: Consider any pair of images

$$I_1(\boldsymbol{x}) = I_0(R_{\theta_1}\boldsymbol{x}), \quad I_2 = I_0(R_{\theta_2}\boldsymbol{x})$$

that are elements of $\mathcal{M}_{\text{rot}}$. Since the set of rotations in $\mathbb{R}^2$ forms a group (called the special orthogonal group $SO(2)$), we have the relation

$$I_2(\boldsymbol{x}) = I_1(R_{\theta_1-\theta_2}\boldsymbol{x}) = I_1(R_{\Delta\theta}\boldsymbol{x}). \tag{6.6}$$

Once again, we denote the locations of the centers of mass of I_1 and I_2 as $\check{\boldsymbol{x}}_1$ and $\check{\boldsymbol{x}}_2$ respectively. Observe that the centers of mass of I_1 and I_2 also obey the relation $\check{\boldsymbol{x}}_2 = R_{\Delta\theta}\check{\boldsymbol{x}}_1$. Hence, we have

$$\|\check{\boldsymbol{x}}_2 - \check{\boldsymbol{x}}_1\|_2 \;=\; \|R_{\Delta\theta}\check{\boldsymbol{x}}_1 - \check{\boldsymbol{x}}_1\|_2 = \left\| \left(\begin{bmatrix} \cos\Delta\theta & -\sin\Delta\theta \\ \sin\Delta\theta & \cos\Delta\theta \end{bmatrix} - \mathbb{I}_{2\times 2} \right) \check{\boldsymbol{x}}_1 \right\|_2 .$$

To establish local isometry, we need to show that the EMD between a pair of images exhibits a linear relationship with the magnitude of the distance in articulation space $\Delta\theta$ in the regime where $\Delta\theta$ is small. In such a regime, we can perform a first-order Taylor series expansion to obtain

$$\begin{aligned} \|\check{\boldsymbol{x}}_2 - \check{\boldsymbol{x}}_1\|_2 \;&\approx\; \left\| \left(\begin{bmatrix} 1 & -\Delta\theta \\ \Delta\theta & 1 \end{bmatrix} - \mathbb{I}_{2\times 2} \right) \check{\boldsymbol{x}}_1 \right\|_2 = \left\| \begin{bmatrix} 0 & -\Delta\theta \\ \Delta\theta & 0 \end{bmatrix} \check{\boldsymbol{x}}_1 \right\|_2 \\ &=\; |\Delta\theta| \; \|\check{\boldsymbol{x}}_1\|_2. \end{aligned}$$

However, the quantity $\|\check{\boldsymbol{x}}_1\|$ represents the distance of the center of mass of image I_1 from the origin, which is constant for images belonging to $\mathcal{M}_{\text{rot}}$. Further, we established in (6.5) that the distance between the centers of a pair of images is upper bounded by a constant times the EMD between the images. Hence, for some constant $\alpha > 0$, we have the following lower bound:

$$EMD(I_1, I_2) \geq \alpha|\Delta\theta|. \tag{6.7}$$

We now prove a similar upper bound on the EMD. By definition, the EMD is calculated by considering the minimum over all feasible flows from I_1 to I_2. Consider the (feasible) flow f corresponding to the bijective mapping between $I_1(\boldsymbol{x}) \doteq I_1(R_{\Delta\theta}\boldsymbol{y})$ and $I_2(\boldsymbol{y})$, i.e.,

$$\gamma(\boldsymbol{x}_i, \boldsymbol{y}_j) = \begin{cases} I_1(\boldsymbol{x}_i), & \boldsymbol{x}_i = R_{\Delta\theta}\boldsymbol{y}_j \\ 0, & \text{otherwise.} \end{cases}$$

For small values of $\Delta\theta$, the magnitude of the displacement of the pixel $\boldsymbol{x}_i$ induced by this flow can be approximated as

$$\|\boldsymbol{x}_i - \boldsymbol{y}_j\|_2 \approx |\Delta\theta| \; \|\boldsymbol{x}_i\|_2,$$

and hence the cost of the flow f can be computed by evaluating the right hand side of (6.4). This quantity provides an upper bound for the EMD between images I_1 and I_2 as follows:

$$\begin{aligned} EMD(I_1, I_2) &\leq \sum_{i,j} \gamma(\boldsymbol{x}_i, \boldsymbol{y}_j) \|\boldsymbol{x}_i - \boldsymbol{y}_j\|_2 \\ &= \sum_i I(\boldsymbol{x}_i)\ |\Delta\theta|\ \|\boldsymbol{x}_i\|_2. \end{aligned}$$

The ℓ_2-norm of $\boldsymbol{x}$ is invariant with respect to rotation, and hence the quantity $\sum_i I(\boldsymbol{x}_j)\ \|\boldsymbol{x}_j\|_2$ is constant across all images I belonging to $\mathcal{M}_{\text{rot}}$. Therefore, for some constant $\beta > 0$, we have the following upper bound:

$$EMD(I_1, I_2) \leq \beta |\Delta\theta|. \tag{6.8}$$

Combining (6.7) and (6.8), we obtain that the manifold $\mathcal{M}_{\text{rot}}$ is approximately isometric to Θ under the EMD metric. □

We numerically illustrate the validity of Proposition 3 in Fig. 6.3. Figure 6.3(a) displays several sample example images formed by rotations of the *Cameraman* test image. As above, we form 100 example pairs of such images, record the distance between the rotation parameter vectors (the "distance in articulation space"), and compute the Euclidean (ℓ_2) distance and EMD between the corresponding images. Figure 6.2(b) clearly indicates that the ℓ_2 distance is largely uninformative with respect to the articulation distance, while the EMD closely correlates with the articulation distance (local isometry).

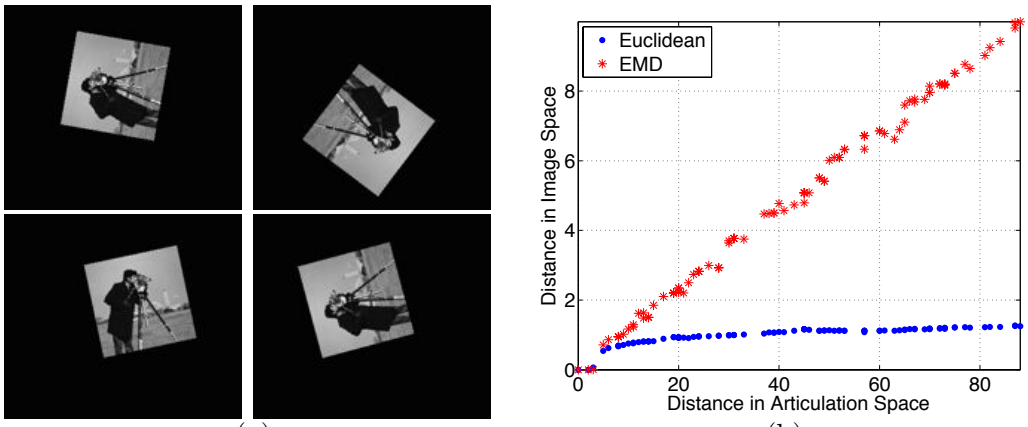
Thus far, we have rigorously proved — for arbitrary translation and rotation manifolds containing images with sharp edges and complex textures — that replacing the ℓ_2 with the EMD surmounts the non-isometry challenge that has plagued the majority of manifold modeling and learning frameworks to date. We now turn to the second challenge of nuisance variables caused by real-world artifacts in the imaging enterprise, such as varying illumination, non-stationary noise and blur, unknown backgrounds, and occlusions.

6.4 Keypoint Articulation Manifolds

Consider the set of images generated by a translating a white in front of a black background under an unknown, spatially varying illumination. Because of the varying illumination, the pixel intensities of the disk will not be constant across the images. In this case, the minimum-cost flow in (6.4) will not be mass-preserving, and the EMD will not be isometric to the translation parameter distance. The standard practical approach to handling illumination variations is to transform the image into a feature-based representation that is robust to such variations. In this section, we propose a systematic framework for analyzing families of articulating images not in terms of their pixel intensities but rather in terms of their *local features).* As we will see, a number of theoretical and practical advantages result.

6.4.1 Feature-based representations for images

We consider local feature representations that consist of a set of image *keypoints* and a corresponding set of *descriptors.* Given an image I defined as a real-valued function over a domain $\Omega \subset \mathbb{R}^2$,



(a) (b)

Figure 6.3: *(a) Sample images from a rotation manifold. (b) Variation of the Euclidean distance and the EMD as a function of the distance in the articulation space. The EMD correlates nearly linearly with articulation distance for the entire range of articulations (local isometry).*

we compute the set of *keypoint locations* $X(I) = \{\boldsymbol{x}_i, i = 1, \ldots, N\} \subset \Omega$ using a local feature extraction algorithm $\mathcal{A}$. At the computed keypoint locations, we compute *keypoint descriptors* $F(I) = \{\boldsymbol{f}_i, i = 1, \ldots, N\}$; each $\boldsymbol{f}_i \in F$ typically can be described as a vector in high-dimensional space $\mathbb{R}^D$. Thus, instead of representing an N-pixel image I as a vector in $\mathbb{R}^N$, we represent it as a set of keypoint location-descriptor pairs $I \sim \{(\boldsymbol{x}_i, \boldsymbol{f}_i), i = 1, \ldots, N\}$, or informally, a "bag of keypoints." Each keypoint location-descriptor pair is an element of an abstract space $\mathcal{X}$ that can be identified with $\mathbb{R}^2 \times \mathbb{R}^D$. Note that $\mathcal{X}$ in itself does not constitute a normed vector space, primarily because the space F is typically not closed under the usual operations of addition and scalar multiplication.

We require that the local feature extraction algorithm $\mathcal{A}$ possess the following properties:
(P1) The keypoint locations are *covariant* to the articulation parameters of interest. For example, in the case of translation, a global translation applied to an image must induce an equivalent, global translation in *every* computed keypoint location.
(P2) The keypoint descriptors are *invariant* to the image articulation parameters of interest.
(P3) The keypoint extraction is *stable*, i.e., no spurious keypoints are detected or missed across different images on the manifold.

Of course, a keypoint extraction algorithm $\mathcal{A}$ exactly satisfying these three properties is hypothetical and may not exist in practice. However, several efficient feature extraction methods have been extensively explored and shown to possess **(P1)**–**(P3)** to a close approximation. The most celebrated is the *Scale Invariant Feature Transform* (SIFT) [150], which approximately possesses **(P1)**–**(P3)** for the case of affine articulations [163] . We will focus on this technique in our computations below without loss of generality.

Definition 8. *Given a keypoint extraction algorithm $\mathcal{A}$ that satisfies properties* **(P1)**–**(P3)** *and an IAM $\mathcal{M} = \{I_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$, the* keypoint articulation manifold *(KAM) is defined as $\mathcal{K} = \{I_{\boldsymbol{\theta}} \sim \{(\boldsymbol{x}_i, \boldsymbol{f}_i)\}_{i=1}^M : I_{\boldsymbol{\theta}} \in \mathcal{M}\}$.*

We seek an appropriate metric on the set K. Consider a grayscale image $I_0(\boldsymbol{x}) \sim \{(\boldsymbol{x}_i, \boldsymbol{f}_i)\}_{i=1}^M$. Define the *keypoint location image* as

$$K_0(\boldsymbol{x}) = \sum_{i=1}^{M} \delta(\boldsymbol{x} - \boldsymbol{x}_i),$$

where $\delta(\cdot)$ is the Kronecker delta function. The keypoint location image can be viewed as a non-negative function over the discrete domain, i.e., $K \in \mathbb{R}_+^N$. Therefore, it is possible to define the EMD between any pair of keypoint location images, which induces a metric on the KAM K. That is, for any pair of images $I_{\boldsymbol{\theta}_1}, I_{\boldsymbol{\theta}_2} \in \mathcal{M}$, we define the *keypoint distance* d_κ as the EMD between their corresponding keypoint location images:

$$d_\kappa(I_{\boldsymbol{\theta}_1}, I_{\boldsymbol{\theta}_2}) = EMD(K_{\boldsymbol{\theta}_1}, K_{\boldsymbol{\theta}_2}).$$

It should be obvious from the properties **(P1)**–**(P3)** that the KAM generated by an ideal keypoint extraction algorithm $\mathcal{A}$ is smooth and globally isometric to any parameter space for which the covariance property **(P1)** holds. We now showcase the power of the invariance property **(P2)**.

6.4.2 Case study: Illumination variations

We prove the following proposition about the geometry of the KAM generated by applying an idealized SIFT-like transformation.

Proposition 4. *Consider an IAM $\mathcal{M}$ generated by images of an arbitrary object as it undergoes 2D translations and in-plane rotations and is then illuminated by an unknown spatially varying illumination. Let K be the KAM generated by applying a keypoint extraction algorithm $\mathcal{A}$ that is covariant to translation and in-plane rotation, and invariant to illumination. Then K, endowed with the keypoint distance d_κ, is globally isometric to the parameter space Θ.*

Proof: We will describe the case where the articulations comprise 2D translations; the extension to in-plane rotations is straightforward and mirrors the derivation in Proposition 3. Any image $I \in \mathcal{M}$ corresponding to the translation parameter $\boldsymbol{\theta}$ can be expressed in terms of a base image I_0 as

$$I(\boldsymbol{x}) = L_{\boldsymbol{\theta}} I_0(\boldsymbol{x} - \boldsymbol{\theta}),$$

where $L_{\boldsymbol{\theta}}$ represents an unknown linear operator representing the illumination corresponding to $\boldsymbol{\theta}$. Consider any pair of images

$$I_1(\boldsymbol{x}) = L_{\boldsymbol{\theta}_1} I_0(\boldsymbol{x} - \boldsymbol{\theta}_1), \quad I_2(\boldsymbol{x}) = L_{\boldsymbol{\theta}_2} I_0(\boldsymbol{x} - \boldsymbol{\theta}_2),$$

that are elements of $\mathcal{M}$. Denote the keypoint location image of I_0 as $K_0(\boldsymbol{x}) = \sum_{i=1}^M \delta(\boldsymbol{x} - \boldsymbol{x}_i)$. By assumption, the algorithm $\mathcal{A}$ stably extracts keypoint locations in a covariant manner, and also is invariant to the illumination operators $L_{\boldsymbol{\theta}_1}, L_{\boldsymbol{\theta}_2}$. Therefore,

$$K_1(\boldsymbol{x}) = K_0(\boldsymbol{x} - \boldsymbol{\theta}_1), \quad K_2(\boldsymbol{x}) = K_0(\boldsymbol{x} - \boldsymbol{\theta}_2),$$

where K_1, K_2 are the corresponding keypoint location images of I_1, I_2. The keypoint distance $d_\kappa(I_1, I_2)$ is equal to the EMD between K_1 and K_2, computed using (6.4). However, in this case the minimum cost flow γ is nothing but a permutation (since K_1, K_2 are the superposition of an identical number M of Kronecker delta functions). Denote $\pi : X(I_{\boldsymbol{\theta}_1}) \to X(I_{\boldsymbol{\theta}_2})$ as a feasible permutation. Therefore,

$$\begin{aligned} EMD(K_1, K_2) &= \min_{\gamma} \sum_{\boldsymbol{x}_i, \boldsymbol{y}_j \in [n]^2} \gamma(\boldsymbol{x}_i, \boldsymbol{y}_j) \|\boldsymbol{x}_i - \boldsymbol{y}_j\|_2 \qquad (6.9) \\ &= \min_{\pi} \sum_{i=1}^{M} \|\boldsymbol{x}_i - \pi(\boldsymbol{x}_i)\|_2. \qquad (6.10) \end{aligned}$$

The optimization (6.10) can be calculated, for example, via the Hungarian algorithm [170]. However, note that, for any permutation π,

$$\begin{aligned} \sum_i \|\boldsymbol{x}_i - \pi(\boldsymbol{x}_i)\| &\geq \left\| \sum_i \boldsymbol{x}_i - \sum_i \pi(\boldsymbol{x}_i) \right\|_2 \\ &= M \left\| \frac{\sum_i \boldsymbol{x}_i}{M} - \frac{\sum_i \pi(\boldsymbol{x}_i)}{M} \right\|_2 \\ &= M \|\check{\boldsymbol{x}}_1 - \check{\boldsymbol{x}}_2\|_2, \end{aligned}$$

where $\check{\boldsymbol{x}}_1, \check{\boldsymbol{x}}_2$ are the centers of mass of the keypoint location images K_1, K_2. Repeating the argument in the proof of Proposition 2, we have that this minimum cost permutation is achieved by mapping the keypoint in K_1 at location $\boldsymbol{x}_i$ to the corresponding keypoint in K_2 at location $\boldsymbol{y}_i \sim \boldsymbol{x}_i + \boldsymbol{\theta}_1 - \boldsymbol{\theta}_2$ (due to the covariance property, this correspondence always exists).

Therefore, $EMD(K_1, K_2)$ is proportional to the distance between the centers of mass of K_1 and K_2, which equals $\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2$. The isometry of the KAM is an immediate consequence. □

6.4.3 Practical computation of the keypoint distance

In order to realize the promise of Proposition 4 in practice, we must address three practical concerns:

1. Noise and numerical errors will render properties **(P1)**–**(P3)** approximations, at best.
2. Real-world phenomena such as occlusions and clutters will also invalidate **(P1)**–**(P3)**. Indeed, accurate detection and filtering of spurious keypoints reduces to establishing exact correspondences between the keypoints, which remains a highly challenging problem in machine vision.
3. The computational complexity of the EMD computation (6.9) is *cubic* in the number of extracted keypoints M, and real-world high-resolution images typically yield several hundreds or even thousands of keypoints [150].

In order to address these challenges, we now propose a computationally efficient approximation to the EMD-based keypoint distance d_κ in (6.9) between any pair of images. We leverage the fact

that the keypoint *descriptors*, $\{\boldsymbol{f}_i\}_{i=1}^{M} \subset \mathbb{R}^D$, calculated from an image $I_{\boldsymbol{\theta}}$ are (approximately) *invariant* to the articulation parameter $\boldsymbol{\theta}$ (recall property **(P2)**). By evaluating a suitably defined *similarity kernel*, $S : \mathbb{R}^D \times \mathbb{R}^D \to \mathbb{R}$, on every pair of keypoint descriptors, we can rapidly establish approximate correspondences between the keypoints. A weighted average of the distances between the corresponding keypoint locations yields the EMD approximation.

The full calculation proceeds as follows. Given a pair of images $I_1 \sim \{(\boldsymbol{x}_i, \boldsymbol{f}_i), i = 1, \ldots, M_1\}$ and $I_2 \sim \{(\boldsymbol{y}_j, \boldsymbol{g}_j), j = 1, \ldots, M_2\}$, we define the *approximate keypoint distance* between I_1 and I_2 as:

$$\begin{aligned} \widetilde{d}_{\kappa}(I_1, I_2) &= \alpha^{-1} \sum_{i,j=1}^{M_1,M_2} S(\boldsymbol{f}_i, \boldsymbol{g}_j) \|\boldsymbol{x}_i - \boldsymbol{y}_j\|_2, \quad \text{where} \\ \alpha &= \sum_{i,j} S(\boldsymbol{f}_i, \boldsymbol{g}_j). \end{aligned} \tag{6.11}$$

The normalization factor α ensures that the approximate keypoint distance does not depend on the *number* of detected keypoint pairs $M_1 \times M_2$. The ideal similarity kernel would yield a value of 1 for every pair of corresponding keypoint locations and zero for all other pairs. In the case when all the keypoint descriptors of the reference image I_0 are distinct, the similarity kernel $S(\boldsymbol{f}_i, \boldsymbol{g}_j)$ would be nonzero only when $\boldsymbol{f}_i \approx \boldsymbol{g}_j$, thereby efficiently *approximating* the minimum cost flow $\gamma(\boldsymbol{x}_i, \boldsymbol{y}_j)$ in (6.9) without an explicit minimization. Consequently, the complexity of evaluating the approximate keypoint distance can be reduced from $\mathcal{O}(M^3)$ to $\mathcal{O}(M^2)$, a significant advantage for practical real-world calculations. We demonstrate this computational advantage numerically in Section 6.5.

The choice of similarity kernel $S(\cdot, \cdot)$ is somewhat flexible. However, to account for numerical discrepancies in the descriptors extracted by the algorithm $\mathcal{A}$, we will focus on the Gaussian radial-basis kernel for $S(\cdot, \cdot)$. For any descriptor pair $(\boldsymbol{f}, \boldsymbol{g})$ and bandwidth parameter $\sigma > 0$, the similarity kernel $S(\cdot, \cdot)$ is given by

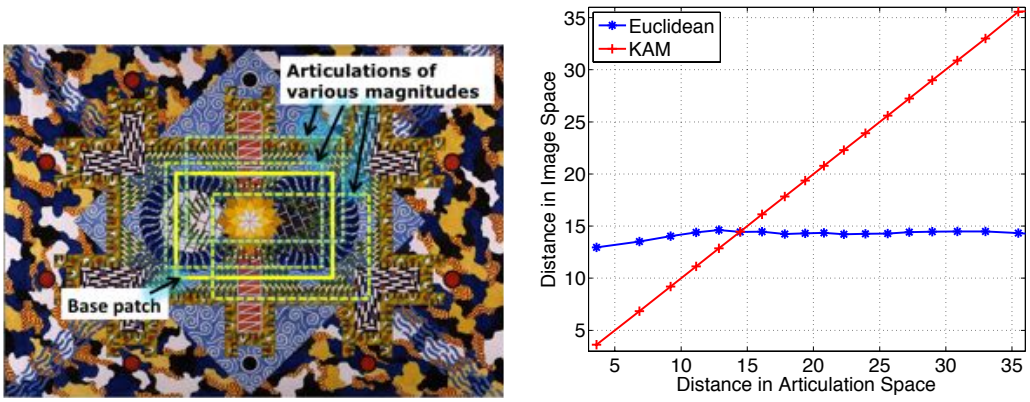
$$S(\boldsymbol{f}, \boldsymbol{g}) = e^{-\frac{\|\boldsymbol{f}-\boldsymbol{g}\|^2}{\sigma^2}}. \tag{6.12}$$

The optimal value of σ in (6.12) depends on the numerical stability of the algorithm $\mathcal{A}$ used to extract feature keypoints from the images. In practice (and for all the experiments below) with SIFT feature keypoints, the value $\sigma = 150$ gave excellent numerical results; moreover, performance is stable to small changes around this value for σ.

Other choices of similarity kernels $S(\cdot, \cdot)$ are also possible. There exist several extensive surveys in the literature on the efficient design of similarity kernels based on local image features [171, 172]. We elaborate on this topic further in Section 6.6.

6.5 Experiments

This experimental section has dual aims. First, we back up the theoretical results on KAM smoothness and isometry using several real-world datasets. Second, we push the KAM technique out of its theoretical comfort zone with new, challenging applications involving a number of real-world



(a) Images (b) Distances

Figure 6.4: *Confirmation of the results of Fig. 6.2 using both the approximate EMD from (6.11) and real data. (a) Sample images from a translation manifold. (b) Variation of the Euclidean distance and the approximate EMD as a function of the distance in the articulation space. The approximate EMD correlates linearly with articulation distance for the entire range of articulations (practically confirming global isometry).*

datasets acquired "in the wild." Our intention is to convincingly demonstrate that manifold methods are not just elegant but also of considerable practical utility in real applications.

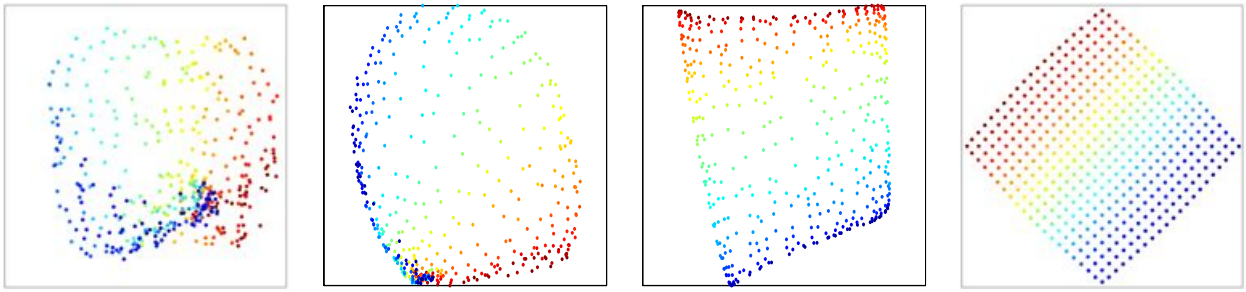
For all experiments that follow, we use the fast approximation to the EMD proposed in (6.11). We use SIFT as the keypoint extraction algorithm $\mathcal{A}$, the Gaussian radial basis function with $\sigma = 150$ for the similarity kernel S, and Isomap (Algorithm 2) for obtaining the low-dimensional embeddings from the distances computed from (6.11). We will refer to this procedure as the "KAM approach".

6.5.1 Confirming smoothness and isometry

Figure 6.4 extends the synthetic experiment in Fig. 6.2 by using both the approximate EMD from (6.11) and real data. We extracted 400 patches of size 80×80 centered at points of a grid of *uniformly-spaced* locations in the highly textured photograph in Fig. 6.4(a)) and replicated the experimental steps of Fig. 6.2. Figure 6.4(b) clearly indicates that the Euclidean (ℓ_2) inter-image distance is largely uninformative with respect to the articulation distance, while the approximate EMD almost perfectly correlates with the articulation distance over the entire range of translations (practically confirming global isometry).

6.5.2 Manifold embedding

We now showcase the invariance and stability properties of the KAM approach with a number of challenging manifold learning (nonlinear dimensionality reduction) examples involving real imagery acquired "in the wild."



(a) Isomap (b) Lap. Eigenmaps (c) Local Features (d) KAM

Figure 6.5: *Continuation of the experiment in Fig. 6.4, plotting 2D embeddings of the translated images using various state-of-the-art manifold learning methods. The KAM approach recovers the underlying parametrization perfectly (modulo a rotation).*

Highly textured translation manifold

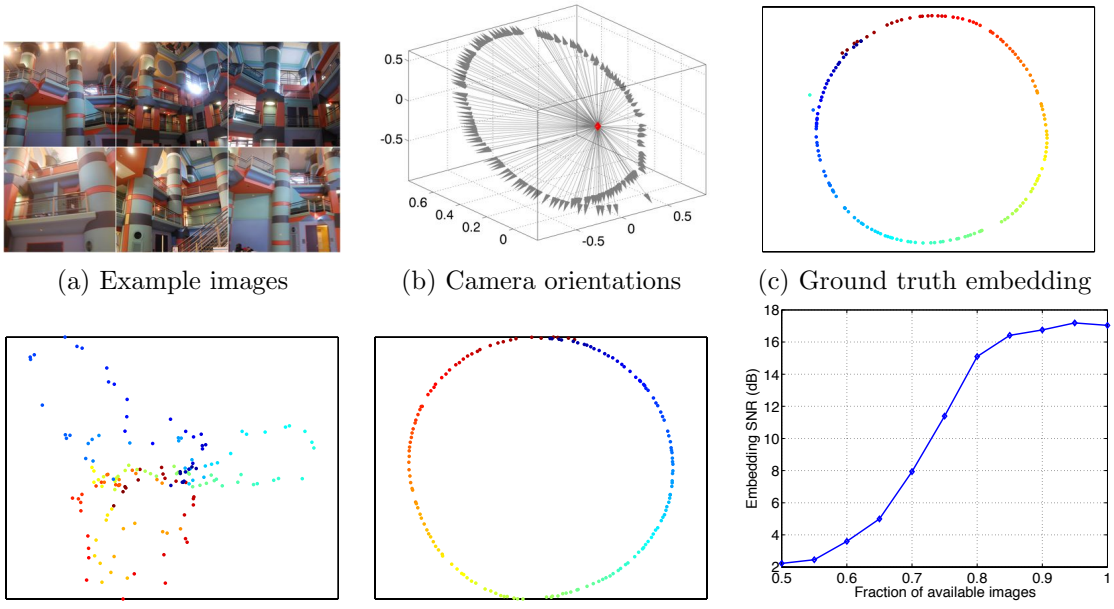
Figure 6.5 continues the example of Fig. 6.4 from Section 6.5.1. Given the sampling of 400 highly textured, translated test images, we ran three state-of-the-art manifold learning algorithms (Isomap, LLE, and Local Features). None of them is able to recover the nonlinear projection into the 2D parameter space as well as our KAM-based Isomap.

Duncan Hall indoor scene

Using a static handheld camera, we collected a set of 160 high-resolution indoor photographs that formed a 360° panoramic view of the walls and ceiling of a large atrium in Rice University's Duncan Hall (see Fig. 6.6). The images are governed not only by an underlying dominant articulation parameter (the viewing angle of the camera), but also by several other degrees of freedom (camera shake and significant lighting variations, including bright sunlight glints). We applied the state-of-the-art structure-from-motion (SfM) Bundler algorithm [159] to estimate, up to an arbitrary rotation, the 3D camera orientation vector for each sample image. We will regard these vectors as the "ground truth" articulation parameters for each image.

Figure 6.6 displays the low-dimensional (2D) embeddings obtained by Isomap using both the classical IAM (using the Euclidean inter-image distance) and the proposed KAM approach (using the approximate EMD inter-image distance). We note that the KAM embedding recovers a near-perfect approximation (modulo a rotation) of the underlying parametrization, whereas the IAM approach yields poor quality results. Figure 6.7 displays additional embeddings produced by four other maninfold learning algorithms, including the Local Features approach [167]. Clearly the KAM approach is much improved over all of these techniques. This demonstrates that the KAM approach is robust to camera jitter and changing lighting conditions.

We now demonstrate that the KAM approach is robust to the sampling of the manifold. Define the *embedding signal-to-noise-ratio (SNR)* as the negative logarithm of the L_2-error of the 2D KAM embedding measured with respect to the ground truth. Figure 6.6(f) shows that the embedding SNR degrades gracefully even when the KAM-based manifold learning algorithm is presented with



(a) Example images (b) Camera orientations (c) Ground truth embedding

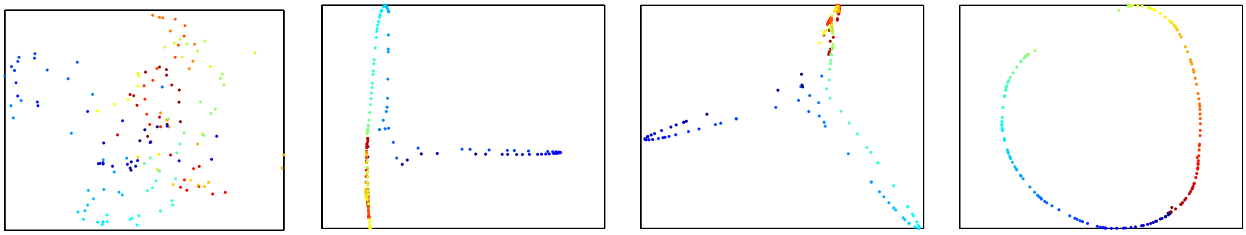
(d) IAM embedding (e) KAM embedding (f) Embedding SNR

Figure 6.6: *Manifold learning in the wild I: Duncan Hall indoor scene. (a) Samples from a set of 160 images obtained from an approximately static hand-held camera that span a 360° panorama of an indoor scene. (b) Camera orientation vectors obtained from the Bundler algorithm to provide precise camera orientation vectors (grey arrows) for each of the images; these can be considered as the "ground truth" values of the underlying parameter space. (c) 2D Isomap embedding of the ground truth camera orientation vectors. (d) 2D Isomap embedding of the IAM using the ℓ_2 metric. (e) 2D KAM embedding; it is virtually equivalent to the optimal embedding using the ground truth (up to an arbitrary rotation). (f) Embedding SNR vs. fraction of available images, indicating that the performance of the KAM approach degrades gracefully with manifold subsampling.*

only a random fraction of the 160 available images.

McNair Hall outdoor scene

We collected a set of 180 images of the front facade of Rice University's McNair Hall by walking with a handheld camera in an approximately straight trajectory; therefore, the underlying parameter space is topologically equivalent to a subset of the real line $\mathbb{R}^1$. Several sample images are shown in Fig. 6.8(a). We used the SfM Bundler software to estimate the camera locations and orientations; the results are displayed in Fig. 6.8(b). As above, we computed low-dimensional embeddings of the images using Isomap on the set of pairwise Euclidean and approximate EMD image distances. The embedding obtained using the KAM approach closely resembles the "ground truth" embedding and successfully recovers the 1D topology of the image dataset.



(a) Diffusion Maps (b) LLE (c) Lap. Eigenmaps (d) Local Features

Figure 6.7: *Manifold learning in the wild I: Duncan Hall indoor scene. Additional results for the dataset in Figure 6.6. (a, b, c) State-of-the-art manifold learning algorithms based on ℓ_2-distances between images perform poorly. (d) The Local Features (LF) approach fares better. However, the KAM approach (Fig. 6.6(e)) still significantly outperforms the Local Features approach in terms of fidelity to the parameter space.*

Brochstein Pavilion outdoor scene

We performed a similar, but more challenging, experiment by collecting a set of 400 images of two adjoining facades of Rice University's Brochstein Pavilion using a hand-held video camera (see Fig. 6.9(a)). In this case as well, the underlying parameter space is topologically equivalent to a subset of $\mathbb{R}^1$ with a kink. However the scene was dynamic; images were captured in different illumination conditions, and several images feature significant occlusions. Due to these non-idealities, the SfM Bundler software exited without converging to a consistent explanation. However, Fig. 6.9 shows that the KAM approach successfully recovers the underlying 1D topological structure of the data set, in contrast to the IAM appraoch. This demonstrates that the KAM approach is robust to some degree of foreground and background clutter.

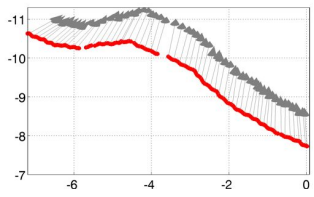
6.5.3 Parameter estimation

We study the effectiveness of the KAM approach for articulation parameter estimation. Given a sample image $I_{\boldsymbol{\theta}} \in \mathcal{M}$, $\boldsymbol{\theta} \in \boldsymbol{\theta}$, our aim is to estimate the underlying vector $\boldsymbol{\theta}$. The non-differentiability of IAMs of images with sharp edges renders IAM-based approaches ineffective for this problem. However, limited progress to date has been made using multiscale smoothing and gradient descent [65]; our goal here is to demonstrate the robust performance of a simple and direct KAM-based estimate.

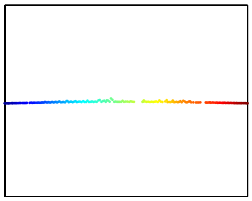
We consider the 400-image translation manifold dataset from Section 6.5.1 and Fig. 6.4 as a "training set". Then, we select a target image patch at random and attempt to estimate its 2D translation parameters by finding the closest among the training set images via a multiscale gradient descent method; the technique used is similar to the method proposed in Section 6.4.1 of [65]. The articulation parameters of the retrieved training image serve as the estimate. We repeat this procedure using both the Euclidean (IAM) and approximate EMD (KAM) distances and record the magnitude of the error between the true and estimated target translation parameters.

Figure 6.10 displays the results of a Monte-Carlo simulation over 40 independent trials. Thanks to the smooth and isometric structure of the KAM, we obtain accurate estimation results even

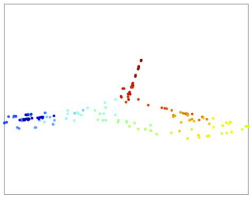
(a) Sample images



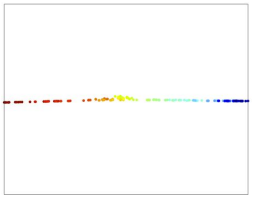
(b) Bundler estimates



(c) Bundler embedding



(d) IAM embedding



(e) KAM embedding

Figure 6.8: *YES WE KAM II: McNair Hall outdoor scene. (a) Samples from a set of 180 images obtained by moving a hand-held camera in an approximately straight trajectory. The image ensemble is topologically equivalent to a 1D manifold. (b) Camera location ground truth obtained from the SfM Bundler algorithm. Camera locations are noted in red and their orientations with grey arrows. (c) 2D Isomap embedding of the ground truth camera orientation vectors. (d) 2D Isomap embedding of the IAM using the ℓ_2 metric. (e) 2D KAM embedding is a close approximation to the ground truth embedding.*

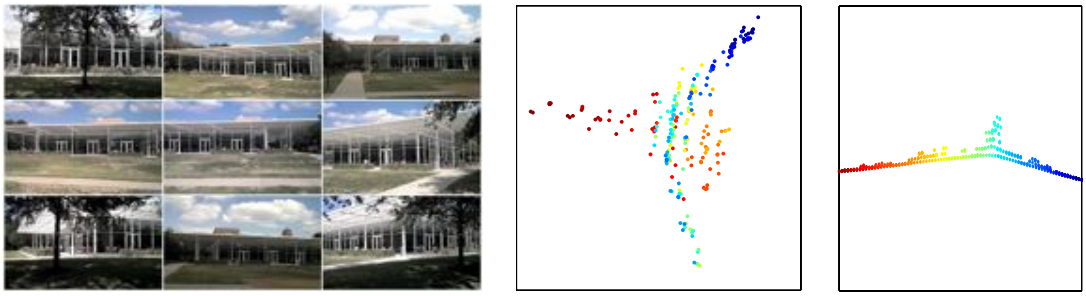
when initializing the gradient descent method far from the target translation value (over 70 pixels, which is significant considering that the images are of size 80×80 pixels). In contrast, the IAM approach suffers from large estimation errors even then starting relatively close to the target value.

We do not claim that this method of estimating the translation parameters via gradient descent on the KAM constitutes a state-of-the-art image registration algorithm. Rather, our aim is merely to show that the smoothness and isometry of the KAM support even naïve information extraction algorithms, in contrast to IAMs.

6.5.4 Organizing photo collections

We now explore how KAMs can be used to automatically organize large collections of images, particularly collections that can be well-modeled by an essentially small number of parameters. An example is the set of photos of a tourist landmark captured by different individuals at different times. The intrinsic variability of this set of photos might be extremely high, owing to occlusions (trees, vehicles, people), variable lighting, and clutter. However, the essential parameters governing the images can be roughly identified with the 3D camera position, orientation, and zoom. We postulate that the KAM approach will help enforce this intrinsic low-dimensionality of the photos and thus provide a meaningful organization. In colloquial terms, we are organizing the photographs by solving a complicated "image jigsaw puzzle" in high-dimensional space by exploiting its low-dimensional geometry

One approach to organize photo collections is the *Photo Tourism* method [159], which runs the SfM Bundler algorithm to accurately estimate the position of each 3D point in the scene and then



(a) Sample images (b) IAM embedding (c) KAM embedding

Figure 6.9: *YES WE KAM III: Brochstein Pavilion outdoor scene. (a) Samples from a set of 400 images obtained by moving a hand-held video camera along two adjoining facades. The image ensemble is topologically equivalent to a 1D manifold with a kink but is complicated by changing illumination and occlusions. (b) 2D Isomap embedding of the ground truth camera orientation vectors. (d) 2D Isomap embedding of the IAM using the ℓ_2 metric. (c) 2D KAM embedding is reflective of the fact that two facades were photographed.*

infers the 3D camera locations and orientations corresponding to each photograph. Unfortunately, while powerful, this algorithm is computationally very demanding and takes several days to execute for a dataset comprising even just a few hundred images.

As an alternative, we propose a far simpler approach: simply extract the keypoints from each of the images, compute the keypoint distances between all pairs of images, and then estimate the geodesics along the KAM. If the low-dimensional manifold assumption holds, then the images corresponding to the nearest neighbors along the geodesics will be semantically meaningful.

Notre Dame

We test our hypothesis on the well-known Notre Dame dataset, a collection of 715 high-resolution images of the popular Parisian tourist trap chosen randomly from FlickR. From each photo, we extract SIFT keypoint locations and descriptors. Using the approximate keypoint distance (6.11), we construct the matrix of pairwise keypoint distances. As in the Isomap algorithm, we use this matrix to construct a $k = 12$-nearest neighbor graph, which we use to estimate the geodesic between any given pair of query images.

Figure 6.11(a) shows the great promise of this proposed technique. We display the seven (geodesic) nearest neighbors for four different query images; it is clear that the retrieved nearest neighbors are closely semantically related to the query image. For comparison purposes, we performed an identical experiment by computing pairwise image distances using the Local Features method [167] and display the results in Fig. 6.11(b). It is evident that the KAM approach results in more semantically meaningful groupings than the Local Features method.

Going one step further, given a pair of starting and ending images, we display the intermediate images along the estimated KAM geodesic in Fig. 6.12. Once again, we observe that the estimated "path" between the photos is both intuitive and interpretable. For example, the images in the bottom row of Fig. 6.12 can be interpreted as zooming out from the inset sculpture to the cathedral

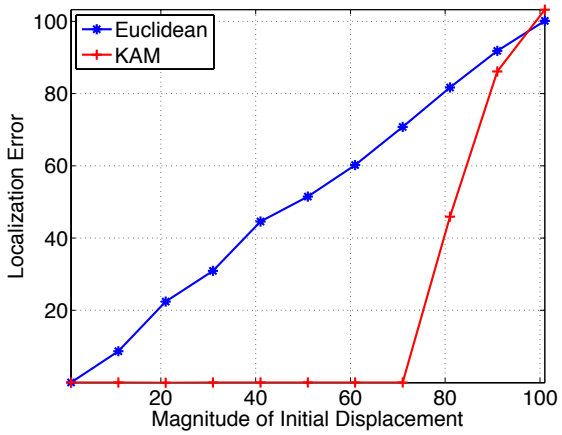


Figure 6.10: *Parameter estimation performance for the translation manifold in Fig. 6.4. The x axis corresponds to the 2D Euclidean distance between the initial translation parameters of the gradient descent and those of the target image. The y axis corresponds to the magnitude of the error between the estimated and target articulations. Gradient descent on the KAM converges accurately for a wide range of initial displacement magnitudes, while gradient descent on the IAM does not yield accurate results for even small values of initial displacement.*

facade. Our method took less than 3 hours to execute in MATLAB.

Statue of Liberty

We repeat the Notre Dame experiment on a database of 2000 images comprising the Statue of Liberty [173] chosen randomly from FlickR. Once again, we extract local image features from each photo and estimate a nearest-neighbor graph using the approximate keypoint distance. Figure 6.13 illustrates that the estimated geodesics between starting and ending images are again semantically meaningful. For example, the images in the top row of Fig. 6.13 can be interpreted as zooming in and panning around the face of the monument.

Of course, our manifold-based method does not produce a full 3D reconstruction of the scene and thus cannot be considered as an alternative to the full 3D modeling technique employed in Photo Tourism [159]. Nevertheless, it can be viewed as a new and efficient way to discover intuitive relationships among photographs. These relationships can potentially be used to improve the performance of algorithms for applications like camera localization and multi-view 3D reconstruction.

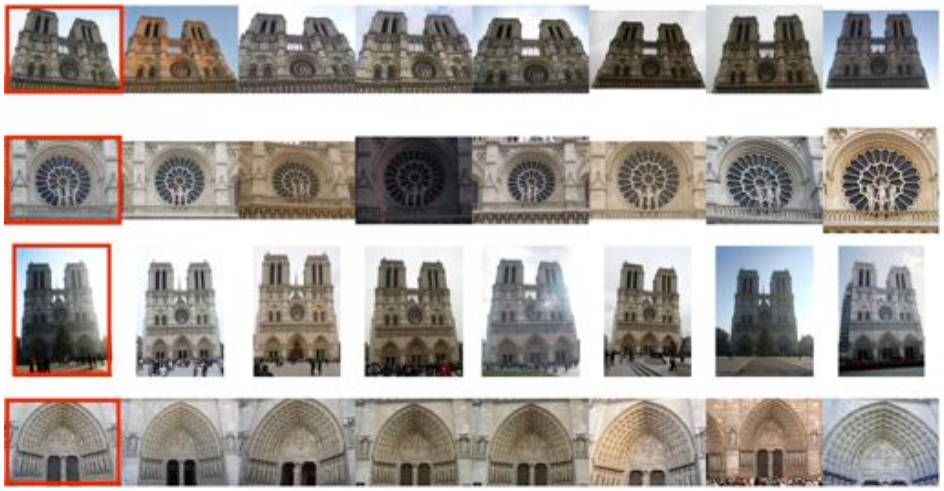
6.6 Discussion

Image manifolds have largely been studied from a theoretical standpoint, and their impact on practical applications has unfortunately not been commensurate to their promise. We have taken some initial steps to bridge this chasm between theory and applications. We have advocated the need for novel distance measures that provide meaningful distances between image pairs and novel

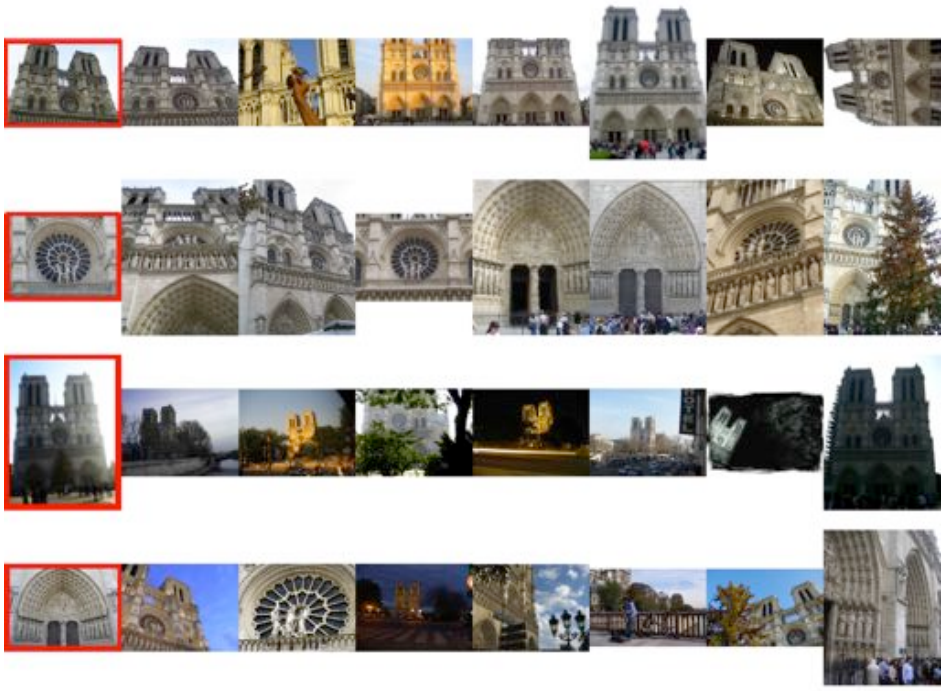
image representations that are robust to nuisance variations. To this end, we have proposed an EMD-based metric on local image features that yields a smooth and isometric mapping from the articulation parameter space to the image feature space.

A first key aspect of our approach is its simplicity. In contrast with the current state-of-the-art methods in SfM, calculating distances in our framework does not involve complicated physics-based modeling of relationships between images, such as epipolar geometry or multi-view stereo. Instead, we merely exploit the low-dimensional manifold geometry inherent in large image ensembles.

A second key aspect of our approach is its computational efficiency. By avoiding explicit correspondence computations between keypoints and image registration, we save significantly on computational complexity. This is reflected in a number of our experiments. The SfM bundler approach [159] greedily establishes correspondences and extracts considerable 3D geometric information from the input images. Yet, it takes several hours, or even days, to produce meaningful results. In contrast, our KAM-based method runs in the order of minutes for data sets of about 150 images and a few hours for a larger dataset of 700+ images. In Chapter 7, we discuss possible directions to extend our approach to even larger datasets.



(a) KAM nearest neighbors



(b) Local Features nearest neighbors

Figure 6.11: *Automatic photo organization using (a) our proposed KAM embedding approach and (b) an approach based on Local Features. The leftmost image in each row (marked in red) indicates the query image, and we retrieve the seven geodesic nearest neighbor images for each query image. In contrast to the Local Features approach, the KAM approach provides more semantically meaningful nearest neighbors.*

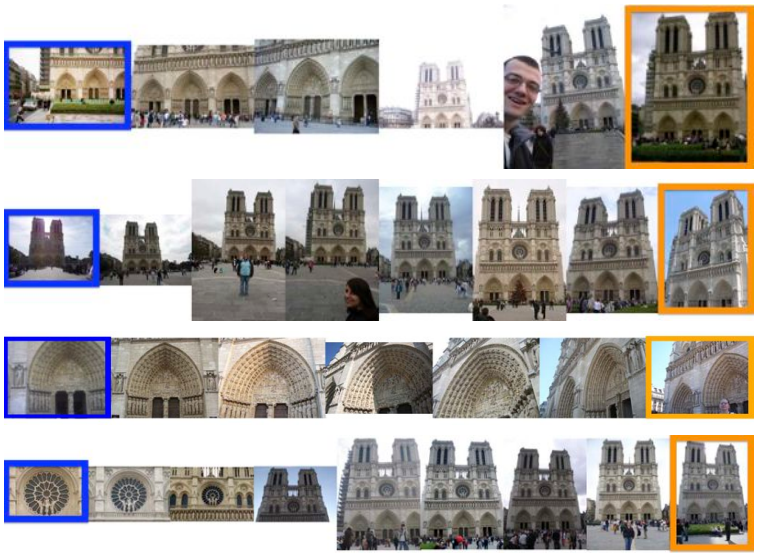


Figure 6.12: *Geodesic paths between images in the Notre Dame dataset. Shown are images along the estimated geodesic for four different choices of start images (marked in blue) and end images (marked in orange).*

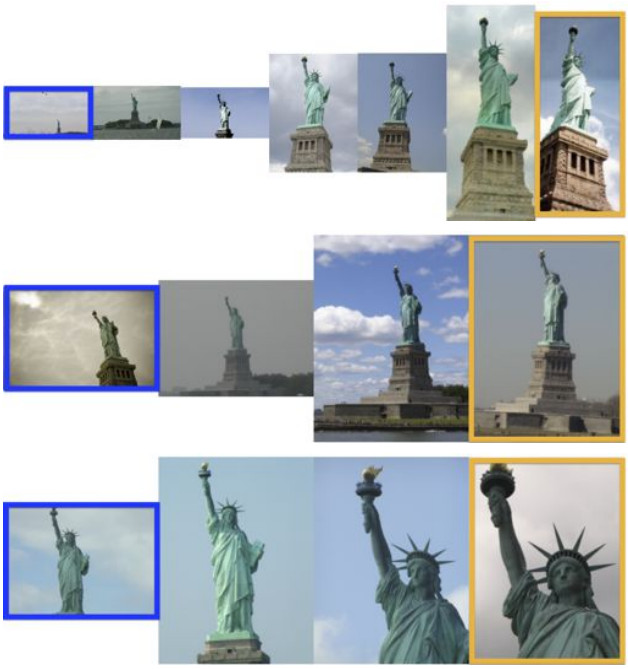


Figure 6.13: *Geodesic paths between images in the Statue of Liberty dataset. Shown are images along the estimated geodesics for three different choices of start images (marked in blue) and end images (marked in orange).*

CHAPTER 7

Conclusions

7.1 Summary

Traditional information systems, based on *linear* modeling principles, continue to bend under the strain of the deluge of data generated, transmitted and processed across the globe. Consequently, there is great need for a unified, comprehensive framework for *nonlinear* signal processing comprising: new, accurate data *models*; new, efficient *algorithms* to process the data; and new *analytical tools* to provide insight and fundamental limits of systems. The ideas developed in this thesis takes some initial steps towards buiding such a framework.

Our mathematical approach is based upon the geometric notion of a nonlinear *manifold* of a higher-dimensional ambient space. Existing models and methods for nonlinear signal acquisition, reconstruction, and inference, can essentially be re-interpreted as specific instances of processing on such manifolds. We have demonstrated that this geometric approach helps unify, analyze, and extend the scope of nonlinear models for a number of important problems encountered in the information processing pipeline.

Our specific contributions have included: a new framework for the design of *linear embeddings* of manifold-modeled data, with applications to signal acquisition, compression, and classification; algorithms and analysis for *signal recovery* from linear samples, with applications to signal reconstruction, deconvolution, and denoising; algorithms and analysis for *efficient inference* from linear samples, with applications to machine learning, parameter estimation, and sensor data fusion; and new representations and algorithms for *robust modeling* and processing of large-scale, real-world image ensembles gathered in the wild.

7.2 New Directions

7.2.1 Semidefinite programming for linear embeddings

We have proposed a scheme for measurement matrix design for manifold-modeled signals and images (Chapter 3). The key idea used in our approach is the *nuclear norm relaxation* of a specific rank minimization problem (3.3). This relaxation paves the way for a tractable optimization (specifically, a semidefinite program (3.4)), thereby enabling the utility of our proposed method for a number of real-world applications.

While our proposed convex relaxation is intuitively justified, a rigorous equivalence between the solution of the rank minimization (3.3) and the optimum achieved by the convex program (3.4) needs to be established. However, note that standard proof techniques for establishing the efficiency of nuclear norm minimization (such as the approach in [58]) are not directly applicable. The main difference is that such techniques heavily rely upon the assumption of randomized, isotropic Gaussian linear measurements of the low-rank matrix of interest, whereas for our problem the linear constraints in (3.4) are cast in a specific rank-1 form.

There exists an alternate approach for establishing the efficiency of the semidefinite relaxation (3.4); specifically, the ideas espoused in [93] are concerned with linear measurements similar to the form posed in (3.4). While the traditional approaches assume a notion of *restricted isometry* on the manifold of low-rank matrices, the approach in [93] advocates the more recent technique of constructing appropriate *matrix dual certificates* to prove the efficiency of the convex relaxation. Suitable modification of the dual certificates argument needs to be performed in order to be applicable to our setting.

7.2.2 Recovery on nonlinear manifolds

In Chapter 4 we proposed and rigorously analyzed an algorithm, which we dub Successive Projections onto INcoherent Manifolds (SPIN), for the recovery of a pair of signals given a small number of measurements of their linear sum. Our proposed algorithm is applicable to a number of common problem settings in signal processing, including deconvolution, denoising, sparse approximation, and Compressive Sensing (CS).

We focus on a particular problem related to *matrix recovery*. The problem of reconstructing, from affine measurements, matrices that are a sum of low-rank and sparse matrices has attracted significant attention in the recent literature [59, 120, 133, 174]. Formally, suppose that we observe measurements:

$$\mathbf{y} = \mathcal{A}(\mathbf{L} + \mathbf{S}),$$

where $\mathbf{L} \in \mathbb{R}^{M \times N}$ is a low-rank matrix, $\mathbf{S} \in \mathbb{R}^{M \times N}$ is a sparse matrix, and $\mathcal{A}$ is a linear measurement operator. The goal is to develop a guaranted, stable algorithm that can recover $(\mathbf{L}, \mathbf{S})$ from $\mathbf{y}$.

An intriguing question is whether our proposed signal recovery algorithm (SPIN) is applicable to this problem. To address this question, we need to verify whether the geometric assumptions, detailed in Section 4.3, are valid for this setting. Certainly, efficient projection operators onto the component manifolds exist. Projection onto the variety of low-rank matrices involves a simple singular value decomposition (SVD), while projection onto the set of sparse matrices involves a

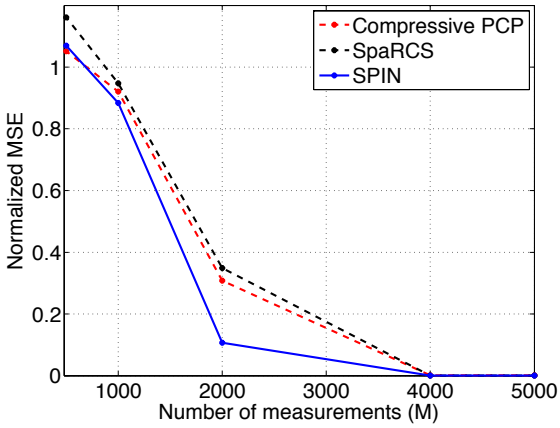


Figure 7.1: *Recovery of low-rank + sparse matrices from compressive measurements. Our proposed algorithm (SPIN) exceeds the performance of SpaRCS, a state-of-the-art recovery algorithm in terms of recovery mean squared error (MSE) for a given number of compressive measurements M.*

simple best K-element approximation. The key stumbling block is that the manifold of low-rank matrices is *not incoherent* with the manifold of sparse matrices. Indeed, the two manifolds share a nontrivial intersection (i.e., there exist low rank matrices that are also sparse, and vice versa). Therefore, guaranteed convergence of SPIN cannot be established as a straightforward consequence of Theorem 2.

Nevertheless, preliminary numerical experiments have demonstrated encouraging numerical results. We consider randomly constructed test matrices $\mathbf{M} = \mathbf{L} + \mathbf{S}$ of size 256×256, where the rank of $\mathbf{L}$ is $r = 3$ and the sparsity of $\mathbf{S}$ is $K = 50$. We form random Gaussian measurements, recover the components $(\mathbf{L}, \mathbf{S})$ using Algorithm 5, and record the recovery error. We discover that the performance of SPIN exceeds that of two state-of-the-art algorithms — SpaRCS [133] and Compressive Principal Component Pursuit [174]; see Fig. 7.1. This experiment, although only an empirical result, is encouraging and it is possible that alternate proof techniques are needed to analyze the superior performance of SPIN for matrix recovery problems.

7.2.3 Robust scalable inference for image ensembles

In Chapter 6 we proposed new image representations and algorithms to enable robust manifold modeling and inference on large-scale image collections gathered "in the wild". The ideas developed in Chapter 6 can be immediately extended to more general settings. For example, the pyramid match kernel [172] is an efficient, robust similarity measure between image pairs that is tailored to object detection. Such a kernel can conceivably be used to induce interesting geometrical structures on IAMs in the same manner as our EMD-inspired similarity kernel (6.12).

We have largely focused on affine articulations in the object or camera, hence motivating our choice of SIFT [150] as the feature extraction algorithm $\mathcal{A}$. A problem involving the manifold of all possible illuminations of an object would likely involve a pose-invariant descriptor. The KAM

approach can be extended to such problems in a principled manner, including proving analytical results along the lines of Propositions 2–4.

We have demonstrated via extensive numerical experiments that the KAM framework offers practical robustness to nuisance phenomena such as background clutter and foreground occlusions. However, modeling such phenomena in a theoretically principled fashion is a difficult task. Particularly challenging scenarios arise in the adversarial setting, where the nuisance clutter and occlusions are deliberately chosen to be perceptually similar to the actual scene of interest. In such a scenario, large and unpredictable errors in the distance computation (6.11) are possible. We leave the precise characterization of the performance of the KAM approach in such circumstances as an open problem.

The primary computational bottleneck in the KAM framework is the calculation of pairwise keypoint distances between images, which scales as $\mathcal{O}\left(M^2\right)$, where M is the number of images. To enable M to scale to tens, or hundreds, of thousands of images or more, the Nyström method [175, 176], which approximates the unknown pairwise distance matrix as low rank and attempts to recover it from a small set of rows and columns of the matrix, can potentially be used Under the same low-rank assumption, a host of techniques from the matrix completion literature [177] can also potentially be applied to recover the pairwise distance matrix from randomly sampled entries. Recently, adaptive selection schemes have been proposed [178] that show improved performance over random selection strategies. All of these schemes can potentially be deployed in conjunction with our proposed framework.

APPENDIX A

Proofs of Chapter 4

A.1 Analysis

The analysis of SPIN is based on the proof technique developed by [122] for analyzing the iterative hard thresholding (IHT) algorithm for sparse recovery and further extended in [96, 123]. For ease of notation, we will assume that $\|\cdot\|$ refers to the Euclidean norm. For a given set of measurements $\mathbf{z}$ obeying (4.1), define the error function $\psi : \mathcal{A} \times B \to \mathbb{R}$ as

$$\psi(\mathbf{a}, \mathbf{b}) = \frac{1}{2} \|\mathbf{z} - \mathbf{\Phi}(\mathbf{a} + \mathbf{b})\|^2 .$$

It is clear that $\psi(\mathbf{a}^*, \mathbf{b}^*) = \frac{1}{2} \|\mathbf{e}\|^2$. The following lemma bounds the error of the estimated signals output by SPIN at the $(k+1)$-st iteration in terms of the error incurred at the k-th iteration, and the norm of the measurement error.

Lemma 8. *Define $(\mathbf{a}_k, \mathbf{b}_k)$ as the intermediate estimates obtained by SPIN at the k-th iteration. Let δ, ϵ be as defined in Theorem 2. Then,*

$$\psi(\mathbf{a}_{k+1}, \mathbf{b}_{k+1}) \leq \alpha \psi(\mathbf{a}_k, \mathbf{b}_k) + C \|\mathbf{e}\|^2 , \tag{A.1}$$

where

$$\alpha = \frac{\frac{2\delta}{1-\delta} + 6\frac{1+\delta}{1-\delta}\frac{\epsilon}{1-\epsilon}}{1 - 4\frac{1+\delta}{1-\delta}\frac{\epsilon}{1-\epsilon}}, \quad C = \frac{\frac{1}{2} + 5\frac{1+\delta}{1-\delta}\frac{\epsilon}{1-\epsilon}}{1 - 4\frac{1+\delta}{1-\delta}\frac{\epsilon}{1-\epsilon}}.$$

Proof. Fix a current estimate of the signal components $(\mathbf{a}_k, \mathbf{b}_k)$ at iteration k. Then, for any other pair of signals $(\mathbf{a}, \mathbf{b}) \in \mathcal{A} \times \mathcal{B}$, we have

$$\begin{aligned}
\psi(\mathbf{a},\mathbf{b}) - \psi(\mathbf{a}_k,\mathbf{b}_k) &= \frac{1}{2}\left(\|\mathbf{z} - \mathbf{\Phi}(\mathbf{a}+\mathbf{b})\|^2 - \|\mathbf{z} - \mathbf{\Phi}(\mathbf{a}_k+\mathbf{b}_k)\|^2\right) \\
&= \frac{1}{2}\left(\|\mathbf{z} - \mathbf{\Phi}(\mathbf{a}+\mathbf{b}) - \mathbf{\Phi}(\mathbf{a}_k+\mathbf{b}_k) + \mathbf{\Phi}(\mathbf{a}_k+\mathbf{b}_k)\|^2 - \|\mathbf{z} - \mathbf{\Phi}(\mathbf{a}_k+\mathbf{b}_k)\|^2\right) \\
&= \frac{1}{2}\left(\|\mathbf{z} - \mathbf{\Phi}(\mathbf{a}_k+\mathbf{b}_k)\|^2 + \|\mathbf{\Phi}(\mathbf{a}_k+\mathbf{b}_k) - \mathbf{\Phi}(\mathbf{a}+\mathbf{b})\|^2 - \|\mathbf{z} - \mathbf{\Phi}(\mathbf{a}_k+\mathbf{b}_k)\|^2\right) \\
&\quad + \langle \mathbf{z} - \mathbf{\Phi}(\mathbf{a}_k+\mathbf{b}_k), \mathbf{\Phi}(\mathbf{a}_k+\mathbf{b}_k) - \mathbf{\Phi}(\mathbf{a}+\mathbf{b})\rangle \\
&= \frac{1}{2}\|\mathbf{\Phi}\mathbf{x} - \mathbf{\Phi}\mathbf{x}_k\|^2 + \langle \mathbf{z} - \mathbf{\Phi}\mathbf{x}_k, \mathbf{\Phi}\mathbf{x}_k - \mathbf{\Phi}\mathbf{x}\rangle,
\end{aligned}$$

where $\mathbf{x}_k \triangleq \mathbf{a}_k + \mathbf{b}_k$, $\mathbf{x} \triangleq \mathbf{a} + \mathbf{b}$. Since $\mathbf{\Phi}$ is a linear operator, we can take the adjoint within the inner product to obtain

$$\begin{aligned}
\psi(\mathbf{a},\mathbf{b}) - \psi(\mathbf{a}_k,\mathbf{b}_k) &= \frac{1}{2}\|\mathbf{\Phi}\mathbf{x} - \mathbf{\Phi}\mathbf{x}_k\|^2 + \langle \mathbf{\Phi}^T(\mathbf{z} - \mathbf{\Phi}\mathbf{x}_k), \mathbf{x}_k - \mathbf{x}\rangle &\text{(A.2)} \\
&\leq \frac{1}{2}(1+\delta)\|\mathbf{x} - \mathbf{x}_k\|^2 + \langle \mathbf{\Phi}^T(\mathbf{z} - \mathbf{\Phi}\mathbf{x}_k), \mathbf{x}_k - \mathbf{x}\rangle. &\text{(A.3)}
\end{aligned}$$

The last inequality occurs due to the RIP of $\mathbf{\Phi}$ applied to the secant vector $\mathbf{x} - \mathbf{x}_k \in \mathcal{S}(\mathcal{C})$. To the right hand side of (A.3), we further add and subtract $\frac{1}{2(1+\delta)}\left\|\mathbf{\Phi}^T(\mathbf{z} - \mathbf{\Phi}\mathbf{x}_k)\right\|^2$ to complete the square:

$$\psi(\mathbf{a},\mathbf{b}) - \psi(\mathbf{a}_k,\mathbf{b}_k) \leq \frac{1}{2}(1+\delta)\left\|\mathbf{x} - \mathbf{x}_k - \frac{1}{1+\delta}\mathbf{\Phi}^T(\mathbf{z} - \mathbf{\Phi}\mathbf{x}_k)\right\|^2 - \frac{1}{2(1+\delta)}\left\|\mathbf{\Phi}^T(\mathbf{z} - \mathbf{\Phi}\mathbf{x}_k)\right\|^2.$$

Define $\mathbf{g}_k \triangleq \frac{1}{1+\delta}\mathbf{\Phi}^T(\mathbf{z} - \mathbf{\Phi}(\mathbf{a}_k + \mathbf{b}_k))$. Then,

$$\psi(\mathbf{a},\mathbf{b}) - \psi(\mathbf{a}_k,\mathbf{b}_k) \leq \frac{1}{2}(1+\delta)\left(\|\mathbf{a}+\mathbf{b} - (\mathbf{a}_k + \mathbf{b}_k + \mathbf{g}_k)\|^2 - \|\mathbf{g}_k\|^2\right). \tag{A.4}$$

Next, define the function ζ on $\mathcal{A} \times \mathcal{B}$ as $\zeta(\mathbf{a},\mathbf{b}) \triangleq \|\mathbf{a}+\mathbf{b} - (\mathbf{a}_k + \mathbf{b}_k + \mathbf{g}_k)\|^2$. Then, we have

$$\begin{aligned}
\zeta(\mathbf{a}_{k+1},\mathbf{b}_{k+1}) &= \|\mathbf{a}_{k+1} - (\mathbf{a}_k + \mathbf{g}_k) + \mathbf{b}_{k+1} - (\mathbf{b}_k + \mathbf{g}_k) + \mathbf{g}_k\|^2 \\
&= \|\mathbf{a}_{k+1} - (\mathbf{a}_k + \mathbf{g}_k)\|^2 + \|\mathbf{b}_{k+1} - (\mathbf{b}_k + \mathbf{g}_k)\|^2 + \|\mathbf{g}_k\|^2 \\
&\quad +2\langle \mathbf{a}_{k+1} - (\mathbf{a}_k + \mathbf{g}_k), \mathbf{b}_{k+1} - (\mathbf{b}_k + \mathbf{g}_k)\rangle + 2\langle \mathbf{g}_k, \mathbf{a}_{k+1} + \mathbf{b}_{k+1} - (\mathbf{a}_k + \mathbf{b}_k + 2\mathbf{g}_k)\rangle.
\end{aligned}$$

But, as specified in Algorithm 5, $\mathbf{a}_{k+1} = \mathcal{P}_{\mathcal{A}}(\mathbf{a}_k + \mathbf{g}_k)$, and hence $\|\mathbf{a}_{k+1} - (\mathbf{a}_k + \mathbf{g}_k)\| \leq \|\mathbf{a} - (\mathbf{a}_k + \mathbf{g}_k)\|$ for any $\mathbf{a} \in \mathcal{A}$. An analogous relation can be formed between $\mathbf{b}_{k+1}$ and $\mathbf{b}^*$. Hence, we have

$$\begin{aligned}
\|\mathbf{a}_{k+1} - (\mathbf{a}_k + \mathbf{g}_k)\| &\leq \|\mathbf{a}^* - (\mathbf{a}_k + \mathbf{g}_k)\| \quad \text{and} \\
\|\mathbf{b}_{k+1} - (\mathbf{b}_k + \mathbf{g}_k)\| &\leq \|\mathbf{b}^* - (\mathbf{b}_k + \mathbf{g}_k)\|.
\end{aligned}$$

Substituting for $(\mathbf{a}_{k+1}, \mathbf{b}_{k+1})$, we obtain

$$\begin{aligned}
\zeta(\mathbf{a}_{k+1}, \mathbf{b}_{k+1}) &\leq \|\mathbf{a}^* - (\mathbf{a}_k + \mathbf{g}_k)\|^2 + \|\mathbf{b}^* - (\mathbf{b}_k + \mathbf{g}_k)\|^2 + \|\mathbf{g}_k\|^2 \\
&\quad + 2\langle \mathbf{a}_{k+1} - (\mathbf{a}_k + \mathbf{g}_k), \mathbf{b}_{k+1} - (\mathbf{b}_k + \mathbf{g}_k)\rangle + 2\langle \mathbf{g}_k, \mathbf{a}_{k+1} + \mathbf{b}_{k+1} - (\mathbf{a}_k + \mathbf{b}_k + 2\mathbf{g}_k)\rangle \\
&= \|\mathbf{a}^* - (\mathbf{a}_k + \mathbf{g}_k)\|^2 + \|\mathbf{b}^* - (\mathbf{b}_k + \mathbf{g}_k)\|^2 + \|\mathbf{g}_k\|^2 \\
&\quad + 2\langle \mathbf{a}^* - (\mathbf{a}_k + \mathbf{g}_k), \mathbf{b}^* - (\mathbf{b}_k + \mathbf{g}_k)\rangle + 2\langle \mathbf{g}_k, \mathbf{a}^* + \mathbf{b}^* - (\mathbf{a}_k + \mathbf{b}_k + 2\mathbf{g}_k)\rangle \\
&\quad + 2\langle \mathbf{a}_{k+1} - (\mathbf{a}_k + \mathbf{g}_k), \mathbf{b}_{k+1} - (\mathbf{b}_k + \mathbf{g}_k)\rangle - 2\langle \mathbf{a}^* - (\mathbf{a}_k + \mathbf{g}_k), \mathbf{b}^* - (\mathbf{b}_k + \mathbf{g}_k)\rangle \\
&\quad + 2\langle \mathbf{g}_k, \mathbf{a}_{k+1} + \mathbf{b}_{k+1} - (\mathbf{a}^* + \mathbf{b}^*)\rangle.
\end{aligned}$$

Completing the squares, we have:

$$\begin{aligned}
\zeta(\mathbf{a}_{k+1}, \mathbf{b}_{k+1}) &\leq \|\mathbf{a}^* + \mathbf{b}^* - (\mathbf{a}_k + \mathbf{b}_k + \mathbf{g}_k)\|^2 + 2\langle \mathbf{a}_{k+1} - \mathbf{a}_k, \mathbf{b}_{k+1} - \mathbf{b}_k\rangle - 2\langle \mathbf{a}^* - \mathbf{a}_k, \mathbf{b}^* - \mathbf{b}_k\rangle \\
&\quad + 2\langle \mathbf{g}_k, -\mathbf{a}_{k+1} + \mathbf{a}_k - \mathbf{b}_{k+1} + \mathbf{b}_k + \mathbf{a}^* - \mathbf{a}_k + \mathbf{b}^* - \mathbf{b}_k + \mathbf{a}_{k+1} + \mathbf{b}_{k+1} - (\mathbf{a}^* + \mathbf{b}^*)\rangle.
\end{aligned}$$

The last term on the right hand side equals zero, and so we obtain

$$\zeta(\mathbf{a}_{k+1}, \mathbf{b}_{k+1}) \leq \zeta(\mathbf{a}^*, \mathbf{b}^*) + 2\langle \mathbf{a}_{k+1} - \mathbf{a}_k, \mathbf{b}_{k+1} - \mathbf{b}_k\rangle - 2\langle \mathbf{a}^* - \mathbf{a}_k, \mathbf{b}^* - \mathbf{b}_k\rangle.$$

Combining this inequality with (A.4), we obtain the series of inequalities

$$\begin{aligned}
\psi(\mathbf{a}_{k+1}, \mathbf{b}_{k+1}) - \psi(\mathbf{a}_k, \mathbf{b}_k) &\leq \frac{1}{2}(1+\delta)\left(\zeta(\mathbf{a}_{k+1}, \mathbf{b}_{k+1}) - \|\mathbf{g}_k\|^2\right) \\
&\leq \overbrace{\frac{1}{2}(1+\delta)\left(\zeta(\mathbf{a}^*, \mathbf{b}^*) - \|\mathbf{g}_k\|^2\right)}^{\mathbb{T}_1} \\
&\quad + \overbrace{(1+\delta)\left(\langle \mathbf{a}_{k+1} - \mathbf{a}_k, \mathbf{b}_{k+1} - \mathbf{b}_k\rangle - \langle \mathbf{a}^* - \mathbf{a}_k, \mathbf{b}^* - \mathbf{b}_k\rangle\right)}^{\mathbb{T}_2} \quad \text{(A.5)} \\
&= \mathbb{T}_1 + \mathbb{T}_2.
\end{aligned}$$

We can further bound the right hand side of (A.5) as follows. First, we expand $\mathbb{T}_1$ to obtain

$$\begin{aligned}
\mathbb{T}_1 &= \frac{1}{2}(1+\delta)\left(\|\mathbf{a}^* + \mathbf{b}^* - (\mathbf{a}_k + \mathbf{b}_k + \mathbf{g}_k)\|^2 - \|\mathbf{g}_k\|^2\right) \\
&= \frac{1}{2}(1+\delta)\left(\|\mathbf{a}^* + \mathbf{b}^* - (\mathbf{a}_k + \mathbf{b}_k)\|^2 - 2\langle \mathbf{g}_k, \mathbf{a}^* + \mathbf{b}^* - (\mathbf{a}_k + \mathbf{b}_k)\rangle\right) \\
&= \frac{1}{2}(1+\delta)\|\mathbf{x}^* - \mathbf{x}_k\|^2 - \langle \mathbf{\Phi}^T(\mathbf{z} - \mathbf{\Phi}\mathbf{x}_k), \mathbf{x}^* - \mathbf{x}_k\rangle.
\end{aligned}$$

Again, $\mathbf{x}^* - \mathbf{x}_k$ is a secant on the direct sum manifold $\mathcal{C}$. By the RIP property of $\mathbf{\Phi}$, we have

$$\begin{aligned}
\mathbb{T}_1 &\leq \frac{1}{2}\frac{1+\delta}{1-\delta}\|\mathbf{\Phi}\mathbf{x}^* - \mathbf{\Phi}\mathbf{x}_k\|^2 + \langle \mathbf{\Phi}^T(\mathbf{z} - \mathbf{\Phi}\mathbf{x}_k), \mathbf{x}_k - \mathbf{x}^*\rangle \\
&= \frac{1}{2}\left(\frac{1-\delta}{1-\delta} + \frac{2\delta}{1-\delta}\right)\|\mathbf{\Phi}\mathbf{x}^* - \mathbf{\Phi}\mathbf{x}_k\|^2 + \langle \mathbf{\Phi}^T(\mathbf{z} - \mathbf{\Phi}\mathbf{x}_k), \mathbf{x}_k - \mathbf{x}^*\rangle \\
&= \frac{1}{2}\|\mathbf{\Phi}\mathbf{x}^* - \mathbf{\Phi}\mathbf{x}_k\|^2 + \langle \mathbf{\Phi}^T(\mathbf{z} - \mathbf{\Phi}\mathbf{x}_k), \mathbf{x}_k - \mathbf{x}^*\rangle + \frac{\delta}{1-\delta}\|\mathbf{z} - \Phi\mathbf{x}_k\|^2.
\end{aligned}$$

By definition, we have that $\psi(\mathbf{a}_k, \mathbf{b}_k) = \frac{1}{2}\|\mathbf{z} - \Phi\mathbf{x}_k\|^2$. Further, we can substitute $\mathbf{a} = \mathbf{a}^*, \mathbf{b} = \mathbf{b}^*$ in (A.2) to obtain

$$\psi(\mathbf{a}^*, \mathbf{b}^*) - \psi(\mathbf{a}_k, \mathbf{b}_k) = \frac{1}{2}\|\mathbf{\Phi x}^* - \mathbf{\Phi x}_k\|^2 + \langle \mathbf{\Phi}^T(\mathbf{z} - \mathbf{\Phi x}_k), \mathbf{x}_k - \mathbf{x}^* \rangle.$$

Therefore, we have the relation

$$\mathbb{T}_1 \leq \psi(\mathbf{a}^*, \mathbf{b}^*) - \psi(\mathbf{a}_k, \mathbf{b}_k) + \frac{2\delta}{1-\delta}\psi(\mathbf{a}_k, \mathbf{b}_k). \tag{A.6}$$

The term $\mathbb{T}_2$ can be bounded using Lemma 5 as follows. We have

$$\begin{aligned} -\langle \mathbf{a}^* - \mathbf{a}_k, \mathbf{b}^* - \mathbf{b}_k \rangle &\leq |\langle \mathbf{a}^* - \mathbf{a}_k, \mathbf{b}^* - \mathbf{b}_k \rangle|, \\ &\leq \frac{\epsilon}{2(1-\epsilon)}\|\mathbf{x}^* - \mathbf{x}_k\|^2. \end{aligned} \tag{A.7}$$

Further, we have

$$\begin{aligned} \langle \mathbf{a}_{k+1} - \mathbf{a}_k, \mathbf{b}_{k+1} - \mathbf{b}_k \rangle &\leq |\langle \mathbf{a}_{k+1} - \mathbf{a}_k, \mathbf{b}_{k+1} - \mathbf{b}_k \rangle| \leq \frac{\epsilon}{2(1-\epsilon)}\|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2 \\ &= \frac{\epsilon}{2(1-\epsilon)}\|(\mathbf{x}_{k+1} - \mathbf{x}^*) - (\mathbf{x}_k - \mathbf{x}^*)\|^2 \\ &= \frac{\epsilon}{2(1-\epsilon)}\left(\|\mathbf{x}_{k+1} - \mathbf{x}^*\|^2 + \|\mathbf{x}_k - \mathbf{x}^*\|^2 - 2\langle \mathbf{x}_{k+1} - \mathbf{x}^*, \mathbf{x}_k - \mathbf{x}^* \rangle\right) \\ &\leq \frac{\epsilon}{2(1-\epsilon)}\left(\|\mathbf{x}_{k+1} - \mathbf{x}^*\|^2 + \|\mathbf{x}_k - \mathbf{x}^*\|^2 + 2|\langle \mathbf{x}_{k+1} - \mathbf{x}^*, \mathbf{x}_k - \mathbf{x}^* \rangle|\right) \\ &\leq \frac{\epsilon}{2(1-\epsilon)}\left(\|\mathbf{x}_{k+1} - \mathbf{x}^*\|^2 + \|\mathbf{x}_k - \mathbf{x}^*\|^2 + 2\|\mathbf{x}_{k+1} - \mathbf{x}^*\|\,\|\mathbf{x}_k - \mathbf{x}^* \rangle\|\right) \\ &\leq \frac{\epsilon}{(1-\epsilon)}\left(\|\mathbf{x}_{k+1} - \mathbf{x}^*\|^2 + \|\mathbf{x}_k - \mathbf{x}^*\|^2\right), \end{aligned} \tag{A.8}$$

where the last two inequalities follow by applying the Cauchy-Schwartz inequality and the AM-GM inequality. Combining (A.7) and (A.8), $\mathbb{T}_2$ can be bounded above as

$$\begin{aligned} \mathbb{T}_2 &\leq (1+\delta)\frac{\epsilon}{1-\epsilon}\left(\frac{3}{2}\|\mathbf{x}^* - \mathbf{x}_k\|^2 + \|\mathbf{x}^* - \mathbf{x}_{k+1}\|^2\right), \\ &\leq \frac{1+\delta}{1-\delta}\frac{\epsilon}{1-\epsilon}\left(\frac{3}{2}\|\mathbf{\Phi x}^* - \mathbf{\Phi x}_k\|^2 + \|\mathbf{\Phi x}^* - \mathbf{\Phi x}_{k+1}\|^2\right). \end{aligned}$$

But,

$$\begin{aligned} \|\mathbf{\Phi x}^* - \mathbf{\Phi x}_k\|^2 &= \|\mathbf{\Phi x}^* - \mathbf{\Phi x}_k + \mathbf{e} - \mathbf{e}\|^2 \\ &\leq 2\|\mathbf{\Phi x}^* - \mathbf{\Phi x}_k + \mathbf{e}\|^2 + 2\|\mathbf{e}\|^2 = 4\psi(\mathbf{a}_k, \mathbf{b}_k) + 4\psi(\mathbf{a}^*, \mathbf{b}^*) \end{aligned}$$

via the same technique used to obtain (A.8). Similarly,

$$\|\mathbf{\Phi x}^* - \mathbf{\Phi x}_{k+1}\|^2 \leq 4\psi(\mathbf{a}_{k+1}, \mathbf{b}_{k+1}) + 4\psi(\mathbf{a}^*, \mathbf{b}^*).$$

Hence, we obtain

$$\begin{aligned}\mathbb{T}_2 &\leq \frac{1+\delta}{1-\delta}\frac{\epsilon}{1-\epsilon}\left(\frac{3}{2}\left(4\psi(\mathbf{a}_k,\mathbf{b}_k)+4\psi(\mathbf{a}^*,\mathbf{b}^*)\right)+4\psi(\mathbf{a}_{k+1},\mathbf{b}_{k+1})+4\psi(\mathbf{a}^*,\mathbf{b}^*)\right)\\ &= \frac{1+\delta}{1-\delta}\frac{2\epsilon}{1-\epsilon}\left(3\psi(\mathbf{a}_k,\mathbf{b}_k)+2\psi(\mathbf{a}_{k+1},\mathbf{b}_{k+1})+5\psi(\mathbf{a}^*,\mathbf{b}^*)\right). \qquad \text{(A.9)}\end{aligned}$$

Combining (A.5), (A.6), and (A.9), we obtain

$$\begin{aligned}\psi(\mathbf{a}_{k+1},\mathbf{b}_{k+1}) &\leq \psi(\mathbf{a}^*,\mathbf{b}^*)+\frac{2\delta}{1-\delta}\psi(\mathbf{a}_k,\mathbf{b}_k)\\ &\quad+\frac{1+\delta}{1-\delta}\frac{2\epsilon}{1-\epsilon}\big(3\psi(\mathbf{a}_k,\mathbf{b}_k)+2\psi(\mathbf{a}_{k+1},\mathbf{b}_{k+1})+5\psi(\mathbf{a}^*,\mathbf{b}^*)\big).\end{aligned}$$

Rearranging, we obtain Lemma 8. □

A.2 Proof of Theorem 2

Equation A.1 describes a linear recurrence relation for the sequence of positive real numbers $\psi(\mathbf{a}_k,\mathbf{b}_k),\ k=0,1,2,\ldots$ with leading coefficient α. By choice of initialization, $\psi(\mathbf{a}_0,\mathbf{b}_0)=\frac{\|\mathbf{z}\|^2}{2}$. Therefore, for all $k\in\mathbb{N}$, we have the relation

$$\begin{aligned}\psi(\mathbf{a}_k,\mathbf{b}_k) &\leq \alpha^k\psi(\mathbf{a}_0,\mathbf{b}_0)+C\frac{1-\alpha^k}{1-\alpha}\|\mathbf{e}\|^2\\ &\leq \alpha^k\psi(\mathbf{a}_0,\mathbf{b}_0)+\frac{C}{1-\alpha}\|\mathbf{e}\|^2.\end{aligned}$$

To ensure that the value of $\psi(\mathbf{a}_k,\mathbf{b}_k)$ does not diverge, the leading coefficient α must be smaller than 1, i.e.,

$$\frac{2\delta}{1-\delta}+6\frac{1+\delta}{1-\delta}\frac{\epsilon}{1-\epsilon}<1-4\frac{1+\delta}{1-\delta}\frac{\epsilon}{1-\epsilon}.$$

Rearranging, we obtain the upper bound on δ as in (4.8):

$$\delta<\frac{1-11\epsilon}{3+7\epsilon}.$$

By choosing $\beta=\frac{C}{1-\alpha}$, and $k\geq T$ such that $T=\lceil\frac{1}{\log(1/\alpha)}\log\frac{\|\mathbf{z}\|^2}{2\nu}\rceil$, Theorem 2 follows.

The proof mechanism of Theorem 3 follows a near-identical procedure as in Lemma 8 and we omit the details for brevity. Also, we observe that Theorem 2 represents merely a sufficient condition for signal recovery; the constants in (4.8) could likely be improved.

APPENDIX B

Proofs of Chapter 5

B.1 Proof of Theorem 5

The proof of Theorem 5 is based on the results describing the effect of the operator $\mathbf{\Phi}$ on the metric structure of nearby points on a given K-dimensional manifold, as described in Section 3.2.4 in [82]. A quick sketch of the proof is as follows. Given a "tolerance" parameter, we can calculate the worst case metric distortion suffered by the manifold under the action of $\mathbf{\Phi}$, such that the estimated correlation dimension of the projected set ΦX is within $(1 + \delta')$ times the correlation dimension of $\mathcal{X}$. We impose a suitable bound on the largest scale used to estimate the correlation dimension (as specified in Equation 2). Finally, we use Theorem 4 to obtain a lower bound in the number of projections required to guarantee the desired accuracy of the ID estimate.

For ease of notation, we will assume that $\|\cdot\|$ refers to the Euclidean norm. Recall that a matrix $\mathbf{\Phi}$ has isometry constant δ over a set $\mathcal{X}$, if for every $\mathbf{x} \in \mathcal{X}$, the following relation holds:

$$(1-\delta)\sqrt{\frac{M}{N}} \leq \frac{\|\mathbf{\Phi x}\|}{\|\mathbf{x}\|} \leq (1+\delta)\sqrt{\frac{M}{N}}. \tag{B.1}$$

We make use of the following lemmata to lead up to the proof.

Lemma 9. *Suppose $r < r_{\max}$, the maximum permissible ball radius used to estimate box counting dimension around any point $\mathbf{x} \in \mathcal{M}$. Let $Q_{\mathbf{x}}(r)$ be the number of points within the ball. If (a) the manifold $\mathcal{M}$ is sampled densely and uniformly, and (b) $2r_{\max} < \tau$ hold true, then the maximum possible value for $Q_{\mathbf{\Phi x}}(r\sqrt{M/N}) = Q_{\mathbf{x}}(r)(1-\delta)^{-\widehat{K}}$, where $\widehat{K}$ is the estimated intrinsic dimension of $\mathcal{M}$.*

Proof. This follows from the assumption that around r, the slope of the graph of $\log Q$ versus $\log r$ is linear with slope $\widehat{K}$. Under projection, owing to Lemma 1, the set of points within the ball

experience an isometry constant δ. Hence, the worst case increase in Q is when every point in a ball of radius $r/(1-\delta)$ gets mapped into a ball of radius $r\sqrt{M/N}$ in the projected space. Hence, by linearity of the graph, the new number of points within the projected ball equals $Q_{\mathbf{x}}(r) \times (1-\delta)^{-\widehat{K}}$. Hence, by a simple union bound argument, the worst case increase in $Q(r)$, obtained by averaging $Q_{\mathbf{x}}(r)$ over all x, is also by a factor of $(1-\delta)^{-\widehat{K}}$. This seems to be a pathological case, but we retain it for our derivation of the greatest possible lower bound on the number of projections. By an identical argument, the worst case decrease in $Q(r)$ is by a factor of $(1+\delta)^{\widehat{K}}$.

Care has to be taken to ensure that $r_{\max}$ is not too large, so as to avoid counting points from "faraway regions" in the manifold. This is captured in assumption (b), which relates $r_{\max}$ to the condition number τ of the manifold. □

Lemma 10. *Suppose $\beta = \log(r_{\max}/r_{\min})$. Then, if $\delta < \beta\delta'/2$, the ID estimate of the projected data set is* guaranteed *to be within δ' times the ID estimate of the original data set.*

Proof. Assuming that the regression is done over $(r_{\min}, r_{\max})$, the slope of the linear region is given simply by $(\ln Q(r_{\max}) - \ln Q(r_{\min}))/\beta$. The worst case, which is again pathological, occurs when the two extremes have been multiplied by $(1+\delta)^{\widehat{K}}$ and $(1-\delta)^{\widehat{K}}$. Therefore, the worst possible error in calculation of the slope is equal to $\widehat{K} \times \log((1+\delta)/(1-\delta))$. Converting to the natural logarithm and applying a Taylor series, we obtain the worst case slope estimation error as $2\widehat{K}\delta\beta$. However, we need this error to be less than $\widehat{K}\delta'$. Rearranging, we get the required upper bound on δ.

Using Lemmas 1 and 2, we are now ready to prove our result. We have shown that in the worst case, δ has to be smaller than the bound given in Lemma 2. Hence eliminating δ from the two equations, we get Equation 2 in Theorem 5 as the constraint. The bound on the number of projections is obtained by simply plugging in the maximum allowable value of δ in Theorem 4. This completes the proof. □

B.2 Proof of Theorem 6

The proof follows in two stages. First, we derive a bound on the errors incurred in estimating pairwise geodesic distances from a suitably constructed graph involving only the randomly projected data points as vertices. Next, given that the errors in estimating geodesic distances are bounded, we derive a bound on the overall residual variance produced by the MDS step in Isomap.

A crucial component in the Isomap algorithm is the calculation of geodesic distances using the graph G. It has been rigorously proved [76] that if the underlying manifold is sampled with sufficiently high density, the estimates of geodesic distances using G well-approximate the lengths of the true underlying geodesics.

B.2.1 Graph distances using random measurements

Let $\mathcal{M}$ and $\{x_i\}$ be as specified in Theorem 6 and suppose G is a graph satisfying the conditions in Main Theorem A of [76]. Note that G is a graph with vertices $\{\mathbf{x}_i\}$ and edges that connect some

(but perhaps not all) pairs of points. The graph distance between two points $\mathbf{u}, \mathbf{v}$ in the set $\{\mathbf{x}_i\}$ is defined as

$$d_G(\mathbf{u}, \mathbf{v}) = \min_P(\|\mathbf{x}_0 - \mathbf{x}_1\| + \cdots + \|\mathbf{x}_{p-1} - \mathbf{x}_p\|),$$

which considers piecewise Euclidean distances over all paths P in the graph G that join $\mathbf{x}_0$ to $\mathbf{v} = \mathbf{x}_p$. Supposing that assumptions 1-7 are met in Main Theorem A (part of the assumption is that only nearby points are connected in G), the conclusion is that

$$(1 - \lambda_1)d_{\mathcal{M}}(\mathbf{u}, \mathbf{v}) \leq d_G(\mathbf{u}, \mathbf{v}) \leq (1 + \lambda_1)d_{\mathcal{M}}(\mathbf{u}, \mathbf{v})$$

for all points $\mathbf{u}, \mathbf{v}$ in the set $\{\mathbf{x}_i\}$. (It claims this holds for all $\mathbf{u}, \mathbf{v} \in \mathcal{M}$, but d_G is defined only for points in G.)

Suppose the data has native dimension N, and let $\mathbf{\Phi} : \mathbb{R}^N \to \mathbb{R}^M$ be a projection operator such that

$$(1 - \beta)\|\mathbf{u} - \mathbf{v}\| \leq \|\mathbf{\Phi u} - \mathbf{\Phi v}\| \leq (1 + \beta)\|\mathbf{u} - \mathbf{v}\|$$

for all $\mathbf{u}, \mathbf{v} \in \mathcal{M}$. (Note that this might be an orthoprojector renormalized by $\sqrt{N/M}$. Also note that this property actually need hold only for all $\mathbf{u}, \mathbf{v} \in \{\mathbf{x}_i\}$.) Suppose also that

$$\frac{1 + \beta}{1 - \beta} < \frac{\epsilon_{\max}}{\epsilon_{\min}}.$$

Now, suppose projections $\{\mathbf{y}_i = \mathbf{\Phi x}_i\}$ of the samples are collected. With these nodes $\{\mathbf{y}_i\}$ as vertices we would like to construct a new graph $\mathbf{\Phi}G$, and the key question is which nodes should be joined by edges. This must be concluded purely from the projections $\{\mathbf{y}_i\}$ themselves, and not using the original connectivity of G as side information. Once the connectivity for $\mathbf{\Phi}G$ is defined, we can define a new distance metric on the points $\mathbf{w}, \mathbf{z} \in \{\mathbf{y}_i\}$:

$$d_{\mathbf{\Phi}G}(\mathbf{w}, \mathbf{z}) = \min_P(\|\mathbf{y}_0 - \mathbf{y}_1\| + \cdots + \|\mathbf{y}_{p-1} - \mathbf{y}_p\|)$$

which considers piecewise Euclidean distances over all paths P in the graph $\mathbf{\Phi}G$ that join $\mathbf{w} = \mathbf{y}_0$ to $\mathbf{z} = \mathbf{y}_p$. Ultimately we hope to conclude that

$$d_{\mathcal{M}}(\mathbf{u}, \mathbf{v}) \approx d_{\mathbf{\Phi}G}(\mathbf{\Phi u}, \mathbf{\Phi v})$$

for all $\mathbf{u}, \mathbf{v} \in \{\mathbf{x}_i\}$.

To define connectivity in $\mathbf{\Phi}G$, our rule is that two nodes $\mathbf{w}, \mathbf{z} \in \{\mathbf{y}_i\}$ should be joined by an edge if and only of

$$\|\mathbf{w} - \mathbf{z}\| \leq (1 + \beta)\epsilon_{\min}.$$

(Actually it is also acceptable to optionally permit any edge of length up to $(1 - \beta)\epsilon_{\max}$, but no greater.) Let us furthermore define a second graph G' on the native data points $\{\mathbf{x}_i\}$ as follows: two nodes $\mathbf{u}, \mathbf{v}$ in G' are connected by an edge if and only if their projections $\mathbf{\Phi u}, \mathbf{\Phi v} \in \{\mathbf{y}_i\}$ are joined by an edge in $\mathbf{\Phi}G$. It is easy to check that G' (like G) meets all assumptions in Main Theorem A. To check assumption 1, suppose $\|\mathbf{u} - \mathbf{v}\| \leq \epsilon_{\min}$. Then $\|\mathbf{\Phi u} - \mathbf{\Phi v}\| \leq (1+\beta)\|\mathbf{u} - \mathbf{v}\| \leq (1+\beta)\epsilon_{\min}$, and so $\mathbf{\Phi u}$ connects to $\mathbf{\Phi v}$ in $\mathbf{\Phi}G$ and hence $\mathbf{u}$ connects to $\mathbf{v}$ in G'. To check assumption 2, suppose

$\|\mathbf{u}-\mathbf{v}\| > \epsilon_{\max}$. Then $\|\mathbf{\Phi u}-\mathbf{\Phi v}\| \geq (1-\beta)\|\mathbf{u}-\mathbf{v}\| > (1-\beta)\epsilon_{\max} > \frac{1+\beta}{1-\beta}\epsilon_{\min} > (1+\beta)\epsilon_{\min}$, and so $\mathbf{\Phi u}$ does not connect to $\mathbf{\Phi v}$ in $\mathbf{\Phi}G$ and hence $\mathbf{u}$ does not connect to $\mathbf{v}$ in G'.

We can see that distances in G' and $\mathbf{\Phi}G$ must be similar. Let let P' be a path in G' joining $\mathbf{u},\mathbf{v} \in \{\mathbf{x}_i\}$ and let $\mathbf{\Phi}P$ be the corresponding path in $\mathbf{\Phi}G$ joining $\mathbf{\Phi u},\mathbf{\Phi v} \in \{\mathbf{y}_i\}$. Then stepping along $\mathbf{\Phi}P$,

$$\begin{aligned}\|\mathbf{y}_0-\mathbf{y}_1\|+\cdots+\|\mathbf{y}_{p-1}-\mathbf{y}_p\| &= \|\mathbf{\Phi x}_0-\mathbf{\Phi x}_1\|+\cdots+\|\mathbf{\Phi x}_{p-1}-\mathbf{\Phi x}_p\| \\ &\leq (1+\beta)(\|\mathbf{x}_0-\mathbf{x}_1\|+\cdots+\|\mathbf{x}_{p-1}-\mathbf{x}_p\|).\end{aligned}$$

This holds for every path, and a similar lower bound holds for every path. It follows that

$$(1-\beta)d_{G'}(\mathbf{u},\mathbf{v}) \leq d_{\mathbf{\Phi}G}(\mathbf{\Phi u},\mathbf{\Phi v}) \leq (1+\beta)d_{G'}(\mathbf{u},\mathbf{v})$$

for all $\mathbf{u},\mathbf{v} \in \{\mathbf{x}_i\}$ and hence that

$$(1-\beta)(1-\lambda_1)d_{\mathcal{M}}(\mathbf{u},\mathbf{v}) \leq d_{\mathbf{\Phi}G}(\mathbf{\Phi u},\mathbf{\Phi v}) \leq (1+\beta)(1-\lambda_2)d_{\mathcal{M}}(\mathbf{u},\mathbf{v})$$

for all $\mathbf{u},\mathbf{v} \in \{\mathbf{x}_i\}$. Thus the projected graph distances (which can be computed purely from the projected data $\{\mathbf{y}_i\}$) provide a faithful approximation of the geodesic distance on the manifold.

B.2.2 Isomap residual variance using perturbed input distances

The final question to be addressed is as follows: how does the performance of Isomap on the data set change, given the maximum perturbation δ that each of our input distances can possibly suffer? We use the residual variance (or stress) as the metric to quantitatively describe the performance of Isomap on a given dataset.

To define stress a little more clearly, let n be the number of sample points and $\mathbf{D} = (d_{rs})^2$ be the $n \times n$ matrix of squared geodesic distances between sample points r and s. Isomap computes the *centered* matrix $\mathbf{B} = (b_{rs})$:

$$b_{rs} = -\frac{d_{rs}^2 - \frac{1}{n}\sum_r d_{rs}^2 - \frac{1}{n}\sum_s d_{rs}^2 + \frac{1}{n^2}\sum_{r,s} d_{rs}^2}{2}.$$

$\mathbf{B}$ is shown to satisfy the relation $\mathbf{B} = \mathbf{X}^T\mathbf{X}$, where $\mathbf{X}$ (size $K \times n$) is the (centered) set of K-dimensional Euclidean coordinates that represents the presumed embedding of the n points. The final step finds the K-dimensional representation of every point by performing an eigenvalue decomposition of $\mathbf{B}$ and obtaining the coordinates by projecting X onto the subspace spanned by the eigenvectors vi corresponding to the K largest eigenvalues $\lambda_i, i = 1,2,\ldots,K$. The stress, or residual variance R, is defined as the sum of the $n-K$ smallest (positive) eigenvalues of $\mathbf{B}$. In an ideal scenario, $\mathbf{B}$ would be of rank K and the smallest $n-K$ eigenvalues (consequently, R) would be equal to zero. R represents the deviation from Euclidean-ness, i.e. the inability of Isomap to embed all distances in Euclidean K-dimensional space.

Now, suppose that d_{rs} is perturbed by a fraction (smaller than δ). We know that an isometry constant of δ implies a squared isometry constant of 3δ. Since we have both upper and lower bounds

on the perturbation of d_{rs}, we can immediately write down the following bound on the distortion suffered by b_{rs}:

$$|\Delta b_{rs}| < 6\delta\Gamma^2,$$

where Γ is the square root of the largest entry of $\mathbf{D}$, or the estimated diameter of our compact manifold. Therefore, under perturbation δ, the matrix B varies as:

$$\mathbf{B}(\delta) = \mathbf{B} + 6\delta\Gamma^2\mathbf{E},$$

where $\mathbf{E}$ is a matrix whose entries are from the interval $(-1, 1)$.

It can be shown [75] that if $\mathbf{B}$ is perturbed by $6\delta\Gamma^2\mathbf{E}$, the eigenvalues λ_i of the new matrix are approximated by the following relation (again, neglecting quadratic terms):

$$\lambda_i(\delta) = \lambda_i + 6\delta\Gamma^2\mathbf{v_i^T C v_i}.$$

Assume that there is a cutoff to distinguish between the K^{th} and the $(K+1)^{\text{th}}$ largest eigenvalues, so that there is no significant reordering of eigenvalues. (A strong way to enforce this would be to impose the condition that δ should be small enough that the first $K+1$ eigenvalues $\lambda_i, i = 1, 2, \ldots, K, K+1$ maintain their respective positions after re-sorting according to absolute value.) Hence, the residual variance as a function of δ can be written as:

$$\begin{aligned} R(\delta) &= \sum_{i=K+1}^{n} \lambda_i(\delta) \\ &= \left(\sum_{i=K+1}^{n} \lambda_i\right) + 6\delta\Gamma^2 \left(\sum_{i=K+1}^{n} \mathbf{v_i^T C v_i}\right) \\ &= R + 6\delta\Gamma^2 \left(\sum_{i=K+1}^{n} \mathbf{v_i^T C v_i}\right). \end{aligned}$$

Since all eigenvectors are orthonormal, the quantity $\mathbf{v}_i^T\mathbf{C}\mathbf{v}_i$ can be bounded by the maximum eigenvalue Λ of the matrix $\mathbf{C}$. Rearranging, we get the following upper bound on the error on the change in the residual variance ΔR:

$$\begin{aligned} \Delta R(\delta) &< 6\delta\Gamma^2\Lambda(n-K) \\ &\approx 6\delta\Gamma^2\Lambda n. \end{aligned}$$

for small K. Therefore, the change in the "average" embedding distortion R_{av} *per sample point*, under the effect of random projections and sampling the manifold, varies with δ as:

$$\Delta R_{\text{av}}(\delta) < 6\delta\Gamma^2\Lambda.$$

B.2.3 Bound on the number of projections

In Section B.2.1, we proved that given a number of random projections of the data, the distances between points calculated by the Isomap algorithm using a suitable connectivity graph

well-approximate the actual geodesic distances. By combining β and $\max(|\lambda_i|)$, we can derive an overall "isometry constant" δ which is guaranteed under the action of the operator $\mathbf{\Phi}$. (If both β and $\max(|\lambda_i|)$ are small, a candidate choice for δ is their sum $(\beta + \max(|\lambda_i|))$.)

The final equation in Section B.2.2 gives us a condition on the number of random projections M required to obtain arbitrarily small δ. This is obtained by plugging the desired value of δ into Theorem 4. Note that Λ and Γ could potentially be large, thus yielding a large prescribed value for the number of projections. Λ is bounded (since $\mathbf{C}$ is a bounded linear operator) and depends only on the size of $\mathbf{C}$. However, this is just a sufficiency condition and in practice, we can make do with far fewer measurements. Also, there is no dependence on N, the ambient dimension, in the bound derived above. Thus, the advantages of analyzing random measurements become evident as N becomes intractably large.

Bibliography

[1] "The New York Times," Dec. 2009. http://bits.blogs.nytimes.com/2009/12/09/the-american-diet-34-gigabytes-a-day/.

[2] A. Szalay and J. Gray, "2020 computing: Science in an exponential world," *Nature*, vol. 440, no. 7083, pp. 413–414, 2006.

[3] R. Baraniuk, "More is less: Signal processing and the data deluge," *Science*, vol. 331, no. 6018, p. 717, 2011.

[4] G. Bell, T. Hey, and A. Szalay, "Beyond the data deluge," *Science*, vol. 323, no. 5919, pp. 1297–1298, 2009.

[5] J. Gantz, "The diverse and exploding digital universe," *IDC Whitepaper*, vol. 2, pp. 1–16, 2008.

[6] J. Mullins, "Ring of steel," *IEEE Spectrum*, vol. 43, no. 7, pp. 12–13, 2006.

[7] C. Shannon, "Communication in the presence of noise," *Proc. Inst. Radio. Eng.*, vol. 37, no. 1, pp. 10–21, 1949.

[8] H. Nyquist, "Certain topics in telegraph transmission theory," *Trans. AIEEE*, vol. 47, no. 2, pp. 617–644, 1928.

[9] E. Whittaker, "On the functions which are represented by the expansions of the interpolation-theory.," *Proc. Royal Soc. Edinburgh, Sec. A*, vol. 35, pp. 181–194, 1915.

[10] V. Kotelnikov, "On the truncation capacity of "ether" and wire in electrocommunications," in *First All-Union Conf. Questions of Comm. (Moscow)*, 1933.

[11] A. Oppenheim and A. Willsky, *Signals and systems*. Prentice-Hall, 1996.

[12] D. Johnson and D. Dudgeon, *Array signal processing: concepts and techniques*. Simon and Schuster, 1992.

[13] H. Kwakernaak and R. Sivan, *Linear optimal control systems*, vol. 172. Wiley-Interscience New York, 1972.

[14] M. Lustig, J. Santos, J. Lee, D. Donoho, and J. Pauly, "Application of compressed sensing for rapid mr imaging," in *Proc. Work. Struc. Parc. Rep. Adap. Signaux (SPARS)*, (Rennes, France), Nov. 2005.

[15] D. MacKay, "Good error-correcting codes based on very sparse matrices," *IEEE Trans. Inform. Theory*, vol. 45, no. 2, pp. 399–431, 1999.

[16] K. Varshney, M. Cetin, J. Fisher, and A. Willsky, "Sparse representation in structured dictionaries with application to synthetic aperture radar," *IEEE Trans. Sig. Proc.*, vol. 56, no. 8, pp. 3548–3561, 2008.

[17] G. Hennenfent and F. Herrmann, "Simply denoise: wavefield reconstruction via jittered undersampling," *Geophysics*, vol. 73, no. 3, pp. 19–28, 2008.

[18] E. Candès and T. Tao, "Near optimal signal recovery from random projections: Universal encoding strategies?," *IEEE Trans. Inform. Theory*, vol. 52, no. 12, pp. 5406–5425, 2006.

[19] E. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. Inform. Theory*, vol. 52, pp. 489–509, Feb. 2006.

[20] E. Candès, "Compressive sampling," in *Proc. Int. Cong. Math.*, vol. 3, (Madrid, Spain), pp. 1433–1452, 2006.

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

[22] N. Srebro and T. Jaakkola, "Weighted low-rank approximations," in *Proc. Int. Conf. Mach. Learning*, (Washington, DC), August 2003.

[23] B. Recht, M. Fazel, and P. Parrilo, "Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization," *SIAM Review*, vol. 52, no. 3, pp. 471–500, 2010.

[24] J. Wright, A. Ganesh, S. Rao, Y. Peng, and Y. Ma, "Robust principal component analysis: Exact recovery of corrupted low-rank matrices by convex optimization," in *Adv. Neural Inf. Proc. Sys.*, December 2009.

[25] M. Fazel, H. Hindi, and S. Boyd, "Rank minimization and applications in system theory," in *American Control Conference, 2004. Proceedings of the 2004*, vol. 4, pp. 3273–3278, IEEE, 2004.

[26] M. Belkin and P. Niyogi, "Semi-supervised learning on Riemannian manifolds," *Machine Learning*, vol. 56, pp. 209–239, 2004.

[27] N. Patwari, A. Hero, and A. Pacholski, "Manifold learning visualization of network traffic data," in *Proc. SIGCOMM Work. Mining Network Data*, (Philadelphia, PA), Aug. 2005.

[28] C. Grimes and D. Donoho, "Image manifolds which are isometric to euclidean space," *J. Math. Imag. and Vision*, vol. 23, no. 1, pp. 5–24, 2005.

[29] M. Jordan, Z. Ghahramani, T. Jaakkola, and L. Saul, "An introduction to variational methods for graphical models," *Machine learning*, vol. 37, no. 2, pp. 183–233, 1999.

[30] J. Whittaker, *Graphical models in applied multivariate statistics*, vol. 16. Wiley New York, 1990.

[31] D. Koller and N. Friedman, *Probabilistic graphical models: principles and techniques*. MIT Press, 2009.

[32] P. Müller and F. Quintana, "Nonparametric Bayesian data analysis," *Stat. Science*, pp. 95–110, 2004.

[33] D. Barry, "Nonparametric Bayesian regression," *Annals Stat.*, pp. 934–953, 1986.

[34] R. Horn and C. Johnson, *Matrix Analysis*. Cambridge Univ. Press, 1990.

[35] N. Young, *An Introduction to Hilbert Space*. Cambridge Univ. Press, 1988.

[36] M. B. Wakin, *The Geometry of Low-Dimensional Signal Models*. PhD thesis, Rice Univ., 2006.

[37] M. Hirsch, *Differential topology*. Springer, 1976.

[38] W. Boothby, *An Introduction to Differentiable Manifolds and Riemannian Geometry*. Academic Press, 2003.

[39] P. Niyogi, S. Smale, and S. Weinberger, "Finding the homology of submanifolds with confidence from random samples," Tech. Rep. TR-2004-08, U. Chicago, 2004.

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

[41] A. Bandeira, E. Dobriban, D. Mixon, and W. Sawin, "Certifying the restricted isometry property is hard," *Arxiv preprint arXiv:1204.1580*, 2012. Preprint.

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

[43] Y. Eldar and M. Mishali, "Robust recovery of signals from a structured union of subspaces," *IEEE Trans. Inform. Theory*, vol. 55, no. 11, pp. 5302–5316, 2009.

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

[45] M. Davenport and M. Wakin, "Analysis of Orthogonal Matching Pursuit using the restricted isometry property," *IEEE Trans. Inform. Theory*, vol. 56, no. 9, pp. 4395–4401, 2010.

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

[47] W. Dai and O. Milenkovic, "Subspace pursuit for compressive sensing signal reconstruction," *IEEE Trans. Inform. Theory*, vol. 55, no. 5, pp. 2230–2249, 2009.

[48] T. Blumensath and M. Davies, "Iterative hard thresholding for compressive sensing," *Appl. Comput. Harmon. Anal.*, vol. 27, no. 3, pp. 265–274, 2009.

[49] V. Cevher, M. Duarte, C. Hegde, and R. Baraniuk, "Sparse signal recovery using Markov Random Fields," in *Adv. Neural Inf. Proc. Sys.*, (Vancouver, Canada), Dec. 2008.

[50] V. Cevher, P. Indyk, C. Hegde, and R. Baraniuk, "Recovery of clustered sparse signals from compressive measurements," in *Int. Conf. on Sampling Theory and Applications (SAMPTA)*, (Marseille, France), May 2009.

[51] C. Hegde, M. Duarte, and V. Cevher, "Compressive sensing recovery of spike trains using a structured sparsity model," in *Workshop on Signal Processing with Adaptive Sparse Structured Representations (SPARS)*, (Saint Malo, France), Apr. 2009.

[52] J. Little and D. O'Shea, *Ideals, varieties, and algorithms: an introduction to computational algebraic geometry and commutative algebra*. Springer, 2007.

[53] V. Chandrasekaran, S. Sanghavi, P. Parrilo, and A. Willsky, "Rank-sparsity incoherence for matrix decomposition," *SIAM J. Optimization*, vol. 21, pp. 572–596, 2011.

[54] E. Candès and B. Recht, "Exact matrix completion via convex optimization," *Found. Comput. Math.*, vol. 9, pp. 717–772, 2008.

[55] E. Candès and Y. Plan, "Matrix completion with noise," *Proceedings of the IEEE*, vol. 98, no. 6, pp. 925–936, 2010.

[56] Z. Lin, M. Chen, L. Wu, and Y. Ma, "The augmented Lagrange multiplier method for exact recovery of corrupted low-rank matrices," *Arxiv preprint arXiv:1009.5055*, 2010.

[57] D. Gross, "Recovering low-rank matrices from few coefficients in any basis," *IEEE Trans. Inform. Theory*, vol. 57, no. 3, pp. 1548–1566, 2011.

[58] M. Fazel, E. Candès, B. Recht, and P. Parrilo, "Compressed sensing and robust recovery of low rank matrices," in *Proc. 40th Asilomar Conf. Signals, Systems and Computers*, (Pacific Grove, CA), Nov. 2008.

[59] V. Chandrasekaran, S. Sanghavi, P. Parrilo, and A. Willsky, "Sparse and low-rank matrix decompositions," in *Proc. Allerton Conf. on Comm., Contr., and Comp.*, (Monticello, IL), Sep. 2009.

[60] J. Tenenbaum, V. Silva, and J. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, pp. 2319–2323, 2000.

[61] S. Roweis and L. Saul, "Nonlinear dimensionality reduction by local linear embedding," *Science*, vol. 290, pp. 2323–2326, 2000.

[62] M. Belkin and P. Niyogi, "Using manifold structure for partially labelled classification," in *Adv. Neural Inf. Proc. Sys.*, vol. 15, MIT Press, 2003.

[63] S. Lafon, Y. Keller, and R. Coifman, "Data fusion and multicue data matching by diffusion maps," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 28, no. 11, pp. 1784–1797, 2006.

[64] D. Donoho and C. Grimes, "Image manifolds which are isometric to Euclidean space," *J. Math. Imaging and Vision*, vol. 23, no. 1, 2005.

[65] M. Wakin, D. Donoho, H. Choi, and R. Baraniuk, "The multiscale structure of non-differentiable image manifolds," in *Proc. SPIE Optics Photonics: Wavelets*, (San Diego, CA), Aug. 2005.

[66] B. Moore, "Principal component analysis in linear systems: Controllability, observability, and model reduction," *IEEE Trans. Automat. Control*, vol. 26, no. 1, pp. 17–32, 1981.

[67] R. Dony and S. Haykin, "Optimally adaptive transform coding," *IEEE Trans. Image Proc.*, vol. 4, no. 10, pp. 1358–1370, 1995.

[68] M. Tipping and C. Bishop, "Probabilistic principal component analysis," *J. Royal Statist. Soc B*, vol. 61, no. 3, pp. 611–622, 1999.

[69] R. Fisher, "The use of multiple measurements in taxonomic problems," *Annals of Human Genetics*, vol. 7, no. 2, pp. 179–188, 1936.

[70] H. Harman, *Modern factor analysis*. U. Chicago Press, 1976.

[71] M. Belkin and P. Niyogi, "Laplacian Eigenmaps for dimensionality reduction and data representation," *Neural Comput.*, vol. 15, no. 6, pp. 1373–1396, 2003.

[72] D. Donoho and C. Grimes, "Hessian Eigenmaps: Locally linear embedding techniques for high dimensional data," *Proc. Natl. Acad. Sci.*, vol. 100, no. 10, pp. 5591–5596, 2003.

[73] K. Q. Weinberger and L. K. Saul, "Unsupervised learning of image manifolds by semidefinite programming," *Intl. J. Comp. Vision*, vol. 70, no. 1, pp. 77–90, 2006.

[74] R. R. Coifman and S. Lafon, "Diffusion maps," *Appl. Comput. Harmon. Anal.*, vol. 21, no. 1, pp. 5–30, 2006.

[75] T. Cox and M. Cox, *Multidimensional Scaling*. Boca Raton, FL: Chapman & Hall / CRC, 1994.

[76] M. Bernstein, V. de Silva, J. Langford, and J. Tenenbaum, "Graph approximations to geodesics on embedded manifolds," tech. rep., Stanford Univ., Dec. 2000.

[77] C. Hegde, A. Sankaranarayanan, and R. Baraniuk, "Near-isometric linear embeddings of manifolds," in *Proc. Stat. Sig. Proc.*, (Ann Arbor, MI), Aug. 2012.

[78] C. Hegde, A. Sankaranarayanan, W. Yin, and R. Baraniuk, "A convex approach for learning near-isometric linear embeddings." In preparation, August 2012.

[79] K. Weinberger and L. Saul, "Unsupervised learning of image manifolds by semidefinite programming," *Intl. J. Comp. Vision*, vol. 70, no. 1, pp. 77–90, 2006.

[80] D. Achlioptas, "Database-friendly random projections," in *Proc. Symp. Principles of Database Systems (PODS)*, (Santa Barbara, CA), May 2001.

[81] W. Johnson and J. Lindenstrauss, "Extensions of Lipschitz mappings into a Hilbert space," in *Proc. Conf. Modern Anal. and Prob.*, (New Haven, CT), Jun. 1982.

[82] R. Baraniuk and M. Wakin, "Random projections of smooth manifolds," *Found. Comput. Math.*, vol. 9, no. 1, pp. 51–77, 2009.

[83] K. Clarkson, "Tighter bounds for random projections of manifolds," in *Proc. Symp. Comp. Geom.*, pp. 39–48, ACM, 2008.

[84] R. Baraniuk, M. Davenport, R. DeVore, and M. Wakin, "A simple proof of the restricted isometry property for random matrices," *Const. Approx.*, vol. 28, no. 3, pp. 253–263, 2008.

[85] M. Grant and S. Boyd, "CVX: Matlab software for disciplined convex programming," Feb. 2009. available online at `http://stanford.edu/~boyd/cvx`.

[86] Z. Wen, *First-order methods for semidefinite programming*. PhD thesis, Columbia University, 2009.

[87] N. Linial, E. London, and Y. Rabinovich, "The geometry of graphs and some of its algorithmic applications," *Combinatorica*, vol. 15, no. 2, pp. 215–245, 1995.

[88] S. Dasgupta and A. Gupta, "An elementary proof of the JL lemma," Tech. Rep. TR-99-006, University of California, Berkeley, 1999.

[89] N. Ailon and B. Chazelle, "The fast johnson-lindenstrauss transform and approximate nearest neighbors," *SIAM J. Computing*, vol. 39, no. 1, pp. 302–322, 2010.

[90] N. Alon, "Problems and results in extremal combinatorics," *Discrete Math.*, vol. 273, no. 1, pp. 31–53, 2003.

[91] D. Broomhead and M. Kirby, "The Whitney Reduction Network: A method for computing autoassociative graphs," *Neural Comput.*, vol. 13, pp. 2595–2616, 2001.

[92] D. Broomhead and M. Kirby, "Dimensionality reduction using secant-based projection methods: The induced dynamics in projected systems," *Nonlinear Dynamics*, vol. 41, no. 1, pp. 47–67, 2005.

[93] E. Candes, T. Strohmer, and V. Voroninski, "PhaseLift: Exact and stable signal recovery from magnitude measurements via convex programming," *Arxiv preprint arXiv:1109.4499*, Sept. 2011.

[94] F. Alizadeh, "Interior-point methods in semidefinite programming with applications to combinatorial optimization," *SIAM J. Optimization*, vol. 5, no. 01, 1995.

[95] E. Candès and B. Recht, "Simple bounds for low-complexity model reconstruction," *Arxiv preprint arXiv:1106.1474*, 2011. Preprint.

[96] R. Meka, P. Jain, and I. Dhillon, "Guaranteed rank minimization via singular value projection," in *Adv. Neural Inf. Proc. Sys.*, (Vancouver, BC), Dec. 2010.

[97] A. Barvinok, "Problems of distance geometry and convex properties of quadratic maps," *Discrete and Comput. Geometry*, vol. 13, no. 1, pp. 189–202, 1995.

[98] P. Moscato, M. Norman, and G. Pataki, "On the rank of extreme matrices in semidefinite programs and the multiplicity of optimal eigenvalues," *Mathematics of Operations Research*, vol. 23, no. 2, pp. 339–358, 1998.

[99] B. Kulis, A. Surendran, and J. Platt, "Fast low-rank semidefinite programming for embedding and clustering," in *Proc. Int. AISTATS Conf.*, 2007.

[100] R. Tütüncü, K. Toh, and M. Todd, "Solving semidefinite-quadratic-linear programs using SDPT3," *Math. Prog.*, vol. 95, no. 2, pp. 189–217, 2003.

[101] I. Polik, "Sedumi 1.3," 2010. http://sedumi.ie.lehigh.edu.

[102] J. Douglas and H. Rachford, "On the numerical solution of heat conduction problems in two and three space variables," *Trans. Amer. Math. Soc*, vol. 82, pp. 421–439, 1956.

[103] S. Ma, D. Goldfarb, and L. Chen, "Fixed point and bregman iterative methods for matrix rank minimization," *Math. Prog.*, vol. 128, no. 1, pp. 321–353, 2011.

[104] N. Halko, P. Martinsson, and J. Tropp, "Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions," *SIAM Rev.*, vol. 53, no. 2, pp. 217–288, 2011.

[105] J. Meijerink and H. van der Vorst, "An iterative solution method for linear systems of which the coefficient matrix is a symmetric M-matrix," *Math. Comp.*, vol. 31, no. 137, pp. 148–162, 1977.

[106] D. Liu and J. Nocedal, "On the limited memory BFGS method for large scale optimization," *Math. Prog.*, vol. 45, no. 1, pp. 503–528, 1989.

[107] S. Boyd and L. Vanderberghe, *Convex Optimization*. Cambridge, England: Cambridge Univ. Press, 2004.

[108] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Trans. Inform. Theory*, vol. 13, no. 1, pp. 21–27, 1967.

[109] S. Arya, D. Mount, N. Netanyahu, R. Silverman, and A. Wu, "An optimal algorithm for approximate nearest neighbor searching fixed dimensions," *J. ACM*, vol. 45, no. 6, pp. 891–923, 1998.

[110] G. Shakhnarovich, T. Darrell, and P. Indyk, *Nearest-Neighbor Methods in Learning and Vision: Theory and Practice*. Cambridge, MA: MIT Press, 2005.

[111] P. Indyk and R. Motwani, "Approximate nearest neighbors: towards removing the curse of dimensionality," in *Proc. ACM Symp. Theory of Comput.*, (New York, NY), pp. 604–613, 1998.

[112] B. Russell, A. Torralba, K. Murphy, and W. Freeman, "Labelme: A database and web-based tool for image annotation," *Int. J. Comput. Vision*, vol. 77, no. 1, pp. 157–173, 2008.

[113] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *Int. J. Comput. Vision*, vol. 42, no. 3, pp. 145–175, 2001.

[114] M. Davenport, M. Duarte, M. Wakin, J. Laska, D. Takhar, K. Kelly, and R. Baraniuk, "The smashed filter for compressive classification and target recognition," in *Proc. IS&T/SPIE Symp. Elec. Imag.: Comp. Imag.*, (San Jose, CA), Jan. 2007.

[115] K. Dungan, C. Austin, J. Nehrbass, and L. Potter, "Civilian vehicle radar data domes," in *Proc. SPIE Radar Sensor Tech.*, Apr. 2010.

[116] C. Hegde and R. Baraniuk, "SPIN : Iterative signal recovery on incoherent manifolds," in *Proc. IEEE Int. Symp. Inform. Theory (ISIT)*, (Cambridge, MA), Jul. 2012.

[117] C. Hegde and R. Baraniuk, "Signal recovery on incoherent manifolds," *Arxiv preprint arXiv:1202.1595*, 2012. Preprint.

[118] M. Elad, J.-L. Starck, P. Querre, and D. Donoho, "Simultaneous cartoon and texture image inpainting using morphological component analysis," *Appl. Comput. Harmon. Anal.*, vol. 19, no. 3, pp. 340–358, 2005.

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

[120] E. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?," *J. ACM*, vol. 58, pp. 1–37, May 2011.

[121] I. Daubechies, M. Defrise, and C. DeMol, "An iterative thresholding algorithm for linear inverse problems with a sparsity constraint," *Comm. Pure Appl. Math.*, vol. 57, no. 11, pp. 1413–1457, 2004.

[122] R. Garg and R. Khandekar, "Gradient descent with sparsification: an iterative algorithm for sparse recovery with restricted isometry property," in *Proc. Int. Conf. Machine Learning*, (Montreal, Canada), Jun. 2009.

[123] P. Shah and V. Chandrasekharan, "Iterative projections for signal identification on manifolds," in *Proc. Allerton Conf. on Comm., Contr., and Comp.*, (Monticello, IL), Sept. 2011.

[124] D. Donoho, A. Maleki, and A. Montanari, "Message passing algorithms for compressed sensing," *Proc. Natl. Acad. Sci.*, vol. 106, no. 45, pp. 18914–18919, 2009.

[125] M. McCoy and J. Tropp, "Sharp recovery bounds for convex deconvolution, with applications," *Arxiv preprint arXiv:1205.1580*, 2012.

[126] M. Elad and A. Bruckstein, "A generalized uncertainty principle and sparse representation in pairs of bases," *IEEE Trans. Inform. Theory*, vol. 48, no. 9, pp. 2558–2567, 2002.

[127] C. Studer, P. Kuppinger, G. Pope, and H. Bölcskei, "Recovery of sparsely corrupted signals," *IEEE Trans. Inform. Theory*, vol. 58, no. 5, pp. 3115–3130, 2012.

[128] A. Feuer and A. Nemirovski, "On sparse representation in pairs of bases," *IEEE Trans. Inform. Theory*, vol. 49, no. 6, pp. 1579–1581, 2003.

[129] R. Gribonval and M. Nielsen, "Sparse representations in unions of bases," *IEEE Trans. Inform. Theory*, vol. 49, no. 12, pp. 3320–3325, 2003.

[130] M. Wakin, D. Donoho, H. Choi, and R. Baraniuk, "High-resolution navigation on non-differentiable image manifolds," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing (ICASSP)*, (Philadelphia, PA), Mar. 2005.

[131] M. Davenport, P. Boufounos, M. Wakin, and R. Baraniuk, "Signal processing with compressive measurements," *IEEE J. Select. Top. Sig. Proc.*, vol. 4, no. 2, pp. 445–460, 2010.

[132] A. Kyrillidis and V. Cevher, "Recipes for hard thresholding methods," Tech. Rep. 175528, EPFL, Oct. 2011.

[133] A. Waters, A. Sankaranarayanan, and R. Baraniuk, "SpaRCS: Recovering low-rank and sparse matrices from compressive measurements," in *Adv. Neural Inf. Proc. Sys.*, (Granada, Spain), Dec. 2011.

[134] C. Hegde, M. Wakin, and R. Baraniuk, "Random projections for manifold learning," in *Adv. Neural Inf. Proc. Sys.*, vol. 20, (Vancouver, BC), pp. 641–648, Dec. 2007.

[135] M. Davenport, C. Hegde, M. Duarte, and R. Baraniuk, "Joint manifolds for data fusion," *IEEE Trans. Image Proc.*, vol. 19, pp. 2580–2594, Oct. 2010.

[136] M. Wakin, J. Laska, M. Duarte, D. Baron, S. Sarvotham, D. Takhar, K. Kelly, and R. Baraniuk, "An architecture for compressive imaging," in *IEEE Conf. Image Proc.*, (Atlanta, GA), Oct. 2006.

[137] S. Kirolos, J. Laska, M. Wakin, M. Duarte, D. Baron, T. Ragheb, Y. Massoud, and R. Baraniuk, "Analog-to-information conversion via random demodulation," in *Proc. IEEE Dallas Circuits and Systems Work. (DCAS)*, (Dallas, TX), Oct. 2006.

[138] D. Donoho, "Compressed sensing," *IEEE Trans. Inform. Theory*, vol. 52, pp. 1289–1306, September 2006.

[139] P. Grassberger and I. Procaccia, "Measuring the strangeness of strange attractors," *Physica D Nonlinear Phenomena*, vol. 9, pp. 189–208, 1983.

[140] J. Theiler, "Statistical precision of dimension estimators," *Physical Review A*, vol. 41, no. 6, pp. 3038–3051, 1990.

[141] F. Camastra, "Data dimensionality estimation methods: a survey," *Pattern Recog.*, vol. 36, pp. 2945–2954, 2003.

[142] J. Costa and A. Hero., "Geodesic entropic graphs for dimension and entropy estimation in manifold learning," *IEEE Trans. Sig. Proc.*, vol. 52, no. 8, pp. 2210–2221, 2004.

[143] E. Levina and P. Bickel, "Maximum likelihood estimation of intrinsic dimension," in *Adv. Neural Inf. Proc. Sys.*, vol. 17, MIT Press, 2005.

[144] C. Hegde, M. Wakin, and R. Baraniuk, "Random projections for manifold learning - proofs and analysis," Tech. Rep. TREE 0710, Rice University, 2007.

[145] Y. Keller, R. Coifman, S. Lafon, and S. Zucker, "Audio-visual group recognition using diffusion maps," *IEEE Trans. Sig. Proc.*, vol. 58, no. 1, pp. 403–413, 2010.

[146] M. Rabbat, J. Haupt, A. Singh, and R. Nowak, "Decentralized compression and predistribution via randomized gossiping," in *Proc. Int. Symp. Inform. Proc. Sensor Net. (IPSN)*, (Nashville, TN), Apr. 2006.

[147] W. Bajwa, J. Haupt, A. Sayeed, and R. Nowak, "Compressive wireless sensing," in *Proc. Int. Symp. Inform. Proc. Sensor Net. (IPSN)*, (Nashville, TN), Apr. 2006.

[148] B. Kégl, "Intrinsic dimension estimation using packing numbers," in *Adv. Neural Inf. Proc. Sys.*, (Vancouver, BC), Dec. 2002.

[149] C. Hegde, A. Sankaranarayanan, and R. Baraniuk, "Learning manifolds in the wild." Preprint, July 2012.

[150] D. Lowe, "Distinctive image features from scale-invariant keypoints," *Intl. J. Comp. Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[151] X. Miao and R. Rao, "Learning the lie groups of visual invariance," *Neural Comput.*, 2007.

[152] A. Srivastava, M. Miller, and U. Grenander, "Ergodic algorithms on special Euclidean groups for ATR," *Systems And Control In The Twenty-first Century*, 1997.

[153] M. Miller and L. Younes, "Group actions, homeomorphisms, and matching: A general framework," *Intl. J. Comp. Vision*, vol. 41, no. 1, pp. 61–84, 2001.

[154] O. Tuzel, F. Porikli, and P. Meer, "Learning on Lie groups for invariant detection and tracking," in *IEEE Conf. Comp. Vision and Pattern Recog*, 2008.

[155] A. Sankaranarayanan, C. Hegde, S. Nagaraj, and R. G. Baraniuk, "Go with the flow: Optical flow-based transport operators for image manifolds," in *Proc. Allerton Conf. on Comm., Contr., and Comp.*, (Monticello, IL), Sept. 2011.

[156] B. Horn and B. Schunck, "Determining optical flow," *Artif. Intel.*, vol. 17, no. 1-3, pp. 185–203, 1981.

[157] J. Sivic and A. Zisserman, "Video Google: A text retrieval approach to object matching in videos," in *IEEE Intl. Conf. Comp. Vision*, 2003.

[158] Y. Furukawa and J. Ponce, "Accurate, dense, and robust multiview stereopsis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 8, pp. 1362–1376, 2010.

[159] N. Snavely, S. Seitz, and R. Szeliski, "Photo tourism: Exploring photo collections in 3D," *ACM Trans. Graph.*, vol. 25, no. 3, pp. 835–846, 2006.

[160] K. Mikolajczyk and C. Schmid, "Scale and affine invariant interest point detectors," *Intl. J. Comp. Vision*, vol. 60, no. 1, pp. 63–86, 2004.

[161] T. Tuytelaars and K. Mikolajczyk, "Local invariant feature detectors: A survey," *Found. Trends in Comp. Graphics and Vision*, vol. 3, no. 3, pp. 177–280, 2008.

[162] A. Witkin, "Scale-space filtering," in *Intl. Joint Conf. Artificial Intell.*, vol. 2, pp. 1019–1922, 1983.

[163] J.-M. Morel and G. Yu, "On the consistency of the SIFT method," Tech. Rep. 26, CMLA, 2008.

[164] C. Wallraven, B. Caputo, and A. Graf, "Recognition with local features: The kernel recipe," in *IEEE Intl. Conf. Comp. Vision*, pp. 257–264, 2003.

[165] J. Zhang, M. Marszałek, S. Lazebnik, and C. Schmid, "Local features and kernels for classification of texture and object categories: A comprehensive study," *Intl. J. Comp. Vision*, vol. 73, no. 2, pp. 213–238, 2007.

[166] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *IEEE Conf. Comp. Vision and Pattern Recog*, 2005.

[167] M. Torki and A. Elgammal, "Putting local features on a manifold," in *IEEE Conf. Comp. Vision and Pattern Recog*, 2010.

[168] Y. Rubner, C. Tomasi, and L. J. Guibas, "The Earth Mover's Distance as a metric for image retrieval," *Intl. J. Comp. Vision*, vol. 40, no. 2, pp. 99–121, 2000.

[169] O. Pele and M. Werman, "A linear time histogram metric for improved SIFT matching," in *Euro. Conf. Comp. Vision*, 2008.

[170] H. Kuhn, "The Hungarian method for the assignment problem," *Naval Research Logistics*, vol. 2, no. 1-2, pp. 83–97, 1955.

[171] S. Lyu, "Mercer kernels for object recognition with local features," in *IEEE Conf. Comp. Vision and Pattern Recog*, 2005.

[172] K. Grauman and T. Darrell, "The pyramid match kernel: Discriminative classification with sets of image features," in *IEEE Intl. Conf. Comp. Vision*, 2005.

[173] X. Li, C. Wu, C. Zach, S. Lazebnik, and J. M. Frahm, "Modeling and recognition of landmark image collections using iconic scene graphs," in *Euro. Conf. Comp. Vision*, 2008.

[174] J. Wright, A. Ganesh, K. Min, and Y. Ma, "Compressive principal component pursuit," *Arxiv preprint arXiv:1202.4596*, 2012.

[175] C. Williams and M. Seeger, "Using the Nyström method to speed up kernel machines," in *Adv. Neural Inf. Proc. Sys.*, 2001.

[176] C. Fowlkes, S. Belongie, F. Chung, and J. Malik, "Spectral grouping using the Nyström method," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 2, pp. 214–225, 2004.

[177] E. Candès and B. Recht, "Exact matrix completion via convex optimization," *Found. Comp. Math.*, vol. 9, no. 6, pp. 717–772, 2009.

[178] B. Eriksson, G. Dasarathy, A. Singh, and R. Nowak, "Active clustering: Robust and efficient hierarchical clustering using adaptively selected similarities," *Arxiv preprint arXiv:1102.3887*, 2011.