JOINT RECOVERY OF HIGH-DIMENSIONAL SIGNALS FROM NOISY AND UNDER-SAMPLED MEASUREMENTS USING FUSION PENALTIES

by

Sunrita Poddar

A thesis submitted in partial fulfillment of the
requirements for the Doctor of Philosophy degree
in Electrical and Computer Engineering
in the Graduate College of
The University of Iowa

December 2018

Thesis Supervisor: Associate Professor Mathews Jacob

Graduate College
The University of Iowa
Iowa City, Iowa

CERTIFICATE OF APPROVAL

PH.D. THESIS

This is to certify that the Ph.D. thesis of

Sunrita Poddar

has been approved by the Examining Committee for the thesis requirement for the Doctor of Philosophy degree in Electrical and Computer Engineering at the December 2018 graduation.

Thesis Committee: _____
Mathews Jacob, Thesis Supervisor

_____
Weiyu Xu

_____
Soura Dasgupta

_____
Sajan Goud Lingala

_____
Bijoy Thattaliyath

ACKNOWLEDGEMENTS

When I had taken the decision to pursue a PhD, I was only partially aware of what would lie ahead and what the experience would be like. I had in fact not been sure whether I was taking the right decision. Thankfully, at the end of my PhD, I can happily say that it was the best decision that I could have taken at the time. The six years that I spent at grad school have been full of mixed experiences. But when I look back and think about the overall journey, I can say with certainty that I have grown tremendously as a researcher and as a person, become aware of my shortcomings, and am working to fix then. For all the valuable lessons learned during this time, I have several people to thank.

I would firstly like to thank my advisor Dr Mathews Jacob. I believe that his guidance was instrumental in developing my ability to think about problems critically and communicate my research work effectively. Moreover, his insistence on having high expectations from me also helped me to develop confidence in my own abilities, and aim big for the future. I would also like to thank my committee members Drs Weiyu Xu, Soura Dasgupta, Bijoy Thattaliyath and Sajan Goud Lingala for spending time to understand my research work, and offering their valuable suggestions.

I am very thankful to Dr Deidra Ansah, Dr Bijoy Thattaliyath and Dr Ravi Ashwath for their help in patient recruitment and data collection at the University hospital. Kori Rich, Autumn Craig and Marla Kleingartner were also extremely helpful and patient when I needed a hand to collect data at our research scanners.

Navigating life as a student here would be extremely hard without Cathy Kern and Dina Blanc. They are always ready to answer my questions, and also go out of their way to help in any way possible.

I would also like to acknowledge the help and encouragement of my current and past lab members, and mention that each one of them has left a mark on me in his/her own way.

I am very thankful to my husband, Arvind, whom I met on the first day of my orientation at grad school. Since then, he has become my closest friend and a pillar of support during the various frustrations of PhD life. We have had many stimulating discussions on topics regarding research and also beyond, which have motivated me to always be curious at heart.

I would like to thank my parents, who have always encouraged me to work hard, aim high and yet remain humble. They have always tried to understand and be supportive of my decisions. My Dadu, who passed away shortly after I started my PhD, had always believed in me more than anyone else, and pushed me to do my best. I am sure that he would have been very proud to see me finish my studies. I would also like to thank my Dida, Mashi and my sister Mitil who are always proud and excited at even my smallest achievements, and wish to see me happy above anything else.

ABSTRACT

The presence of missing entries pose a hindrance to data analysis and interpretation. The missing entries may occur due to a variety of reasons such as sensor malfunction, limited acquisition time or unavailability of information. In this thesis, we present algorithms to analyze and complete data which contain several missing entries. We consider the recovery of a group of signals, given a few under-sampled and noisy measurements of each signal. This involves solving ill-posed inverse problems, since the number of available measurements are considerably fewer than the dimensionality of the signal that we aim to recover. In this work, we consider different data models to enable joint recovery of the signals from their measurements, as opposed to the independent recovery of each signal. This prior knowledge makes the inverse problems well-posed. While compressive sensing techniques have been proposed for low-rank or sparse models, such techniques have not been studied to the same extent for other models such as data appearing in clusters or lying on a low-dimensional manifold. In this work, we consider several data models arising in different applications, and present some theoretical guarantees for the joint reconstruction of the signals from few measurements. Our proposed techniques make use of fusion penalties, which are regularizers that promote solutions with similarity between certain pairs of signals.

The first model that we consider is that of points lying on a low-dimensional manifold, embedded in high dimensional ambient space. This model is apt for describing a collection of signals, each of which is a function of only a few parameters; the manifold dimension is equal to the number of parameters. We propose a technique to recover a series of such signals, given a few measurements for each signal. We demonstrate this in the context of dynamic Magnetic Resonance Imaging (MRI) reconstruction, where only a few Fourier measurements are available for each time frame. A novel acquisition scheme enables us to detect the neighbours of each frame on the manifold.

We then recover each frame by enforcing similarity with its neighbours. The proposed scheme is used to enable fast free-breathing cardiac and speech MRI scans.

Next, we consider the recovery of curves/surfaces from few sampled points. We model the curves as the zero-level set of a trigonometric polynomial, whose bandwidth controls the complexity of the curve. We present theoretical results for the minimum number of samples required to uniquely identify the curve. We show that the null-space vectors of high dimensional feature maps of these points can be used to recover the curve. The method is demonstrated on the recovery of the structure of DNA filaments from a few clicked points. This idea is then extended to recover data lying on a high-dimensional surface from few measurements. The formulated algorithm has similarities to our algorithm for recovering points on a manifold. Hence, we apply the above ideas to the cardiac MRI reconstruction problem, and are able to show better image quality with reduced computational complexity.

Finally, we consider the case where the data is organized into clusters. The goal is to recover the true clustering of the data, even when a few features of each data point is unknown. We propose a fusion-penalty based optimization problem to cluster data reliably in the presence of missing entries, and present theoretical guarantees for successful recovery of the correct clusters. We next propose a computationally efficient algorithm to solve a relaxation of this problem. We demonstrate that our algorithm reliably recovers the true clusters in the presence of large fractions of missing entries on simulated and real datasets.

This work thus results in several theoretical insights and solutions to different practical problems which involve reconstructing and analyzing data with missing entries. The fusion penalties that are used in each of the above models are obtained directly as a result of model assumptions. The proposed algorithms show very promising results on several real datasets, and we believe that they are general enough to be easily extended to several other practical applications.

PUBLIC ABSTRACT

Large datasets often contain a wealth of information, and it is the task of data analysis algorithms to discover patterns in this data and make useful inferences from them. Such algorithms are now found abundantly, for various different applications. However, many of these algorithms cannot handle situations when a part of the data is corrupted or missing. A simple example is a survey response where the respondent has chosen to not answer certain questions. Another example is that of satellite data, when images on certain days have obstructions due to cloud cover. Netflix is also a good example of this situation where a large database of movies exists, yet each user is only able to rate a tiny fraction of these. In all the above examples, the cause of missing information is different. Yet, they all create problems for traditional data analysis tools. One aim of this work is to develop techniques to recover the data which has corrupted measurements, i.e. to fill in the missing or corrupted measurements. We develop theory which describes the situations under which the missing entries may be reliably recovered. We also develop some tools to detect patterns in data in the presence of missing entries. The common link between all our developed algorithms is the use of 'fusion penalties' which fills in the missing entries of a particular signal, by searching for other signals that are similar to it.

An important application that we consider is Magnetic Resonance Imaging (MRI). This is a very popular medical imaging modality to study the structure and function of different body parts. We look at dynamic applications such as speech and cardiac imaging, where the aim is to capture the motion of these organs as a function of time. In order to accurately capture the motion, we need to acquire a large number of time frames in a short time. Since, MRI is a very slow imaging modality, it is possible to only partially acquire the samples for each image frame. This results in many missing entries which need to be filled in before the images can be analyzed. The current clinical practice to make the problem less challenging is to ask to patient to hold

his/her breath during each acquisition. We were able to come up with a technique to perform the MRI scan in the free-breathing mode, followed by estimating the missing samples. This is very helpful for critically ill patients who are unable to sustain long breath-holds. We have tested our scheme on several patients in our University Hospital.

Another application that we consider is the estimation of the structure of DNA strands from a few points manually clicked in very poor quality and noisy images. The method can be applied to other problems where we have only a few points in 2D or 3D space, and we want to estimate the underlying curve and surface respectively. We also consider the problem of finding clusters within datasets. An example application is to detect groups of people who have similar interests from some personal information we have regarding each of them. The problem of missing entries is very significant here, since we may not have all the information regarding each person. We apply our proposed technique to the classification of Wine datasets and words from an Australian Sign Language dataset, where have only partial information regarding each data point. We demonstrate that we are able to find accurate clusters even in the presence of these missing entries.

We thus present algorithms for use in a wide variety of applications where missing data is encountered. Our presented algorithms are quite general, and we believe that they can be extended for use in other applications that have not been considered here, which require the estimation of missing entries.

TABLE OF CONTENTS

LIST OF TABLES

Table

LIST OF FIGURES

Figure

CHAPTER 1
INTRODUCTION

1.1 Overview

The problem of data recovery from under-sampled measurements has received much attention in compressive sensing literature. However, the data models considered have mostly been sparse [21] or low rank [11]. Alternate models such as data lying on a low-dimensional manifold or data drawn from multiple clusters have not been studied to the same extent. We propose to develop algorithms for data recovery and analysis, considering such alternate models, motivated by real-world datasets where sparsity or low-rank assumptions do not hold.

We assume that we have a group of signals satisfying a particular data model. An example that we consider here is that the signals lie on a low-dimensional manifold, embedded in high dimensional space. Given some under-sampled measurements of each signal, we study the problem of jointly recovering these signals. Let there be k such signals $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k$. We have the following under-sampled and noisy measurements $\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_k$:

$$\begin{aligned} \mathcal{A}_1(\mathbf{x}_1) &= \mathbf{b}_1 \\ \mathcal{A}_2(\mathbf{x}_2) &= \mathbf{b}_2 \\ &\cdots \\ \mathcal{A}_k(\mathbf{x}_k) &= \mathbf{b}_k \end{aligned} \tag{1.1}$$

where $\{\mathcal{A}_i\}$ are the known measurement operators. The independent recovery of $\mathbf{x}_i$ from its corresponding measurements $\mathbf{b}_i$ is an ill-posed problem. We thus constrain our solutions to the model satisfied by the group of signals $\{\mathbf{x}_i\}$, and recover them jointly from the measurements $\{\mathbf{b}_i\}$. We design regularizers which impose our assumed data model. In this work, we consider fusion penalties, which encourage

solutions with similarity between certain pairs of the recovered signals. Thus, the redudancies present in the dataset are exploited using these penalties. The form of these fusion penalties vary depending on the assumed data model.

We next present the general idea behind joint recovery of signals using fusion penalties. This is followed by some background on the different data models that we consider, as well as a brief discussion on the Magnetic Resonance Imaging (MRI) reconstruction problem, which is one of our important applications.

1.1.1 Joint recovery of signals using fusion penalties

We consider the problem of jointly recovering the signals by solving the following optimization problem:

$$\{\mathbf{x}_i^*\} = \arg\min_{\{\mathbf{x}_i\}} \sum_i \|\mathcal{A}_i(\mathbf{x}_i) - \mathbf{b}_i\|^2 + \lambda \sum_{i,j\in\mathcal{S}} \phi_{i,j}(\|\mathbf{x}_i - \mathbf{x}_j\|_p) \tag{1.2}$$

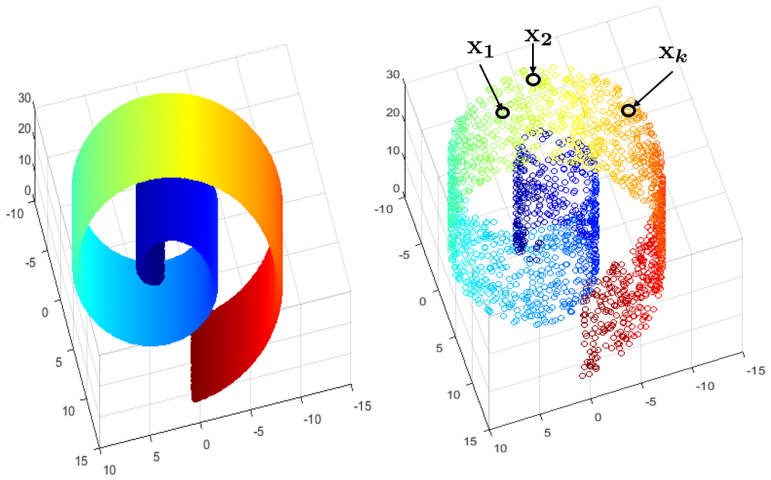
The first term is the data consistency term which imposes an agreement between the recovered signals and our measurements. The second term is a regularizer which imposes pairwise similarity between the recovered signals. This is termed the 'fusion penalty', since it encourages certain recovered signals to 'fuse' and become similar. This behaviour is controlled by the positive non-decreasing functions $\{\phi_{i,j}\}$. In our work, these functions are chosen depending on the assumed data model. The 'p' value in the argument of $\phi_{i,j}$ is usually chosen as 1 or 2. The regularization parameter λ controls the relative importance between the two terms.

One of the earliest applications of the fusion penalty was in [43], where $\phi_{i,j} = \mathcal{I}$, and $\mathcal{S}$ was chosen to contain temporal neighbours. This was later extended in [87] with the introduction of the fused lasso, which enforced sparsity of the recovered signal as well as similarity between successive elements. Later works have also used these penalties for solving inverse problems for signals satisfying the sparsity assumption [27, 35, 50]. However, most existing works on solving inverse problems using fusion

penalties impose similarity only between temporal neighbours, thus disregarding any non-local structure that may be present in the data. Moreover, the functions $\phi_{i,j} = \mathcal{I}$ are not able to exploit any non-linear behaviour in the data. In our work, we we consider models where it is helpful to take into account the non-local redundancies and non-linear behaviour of the data.

1.1.2 Points lying on a low-dimensional manifold

Real data lying in high-dimensional space can often be expressed in terms of only a few parameters. An example is a dataset of face images of the same person with varying pose and illumination. Each image is thus a function of a very low-dimensional parameter vector. Such data points lie on a low-dimensional manifold (with dimension equal to the length of the parameter vector) embedded in high-dimensional space. This is illustrated in Fig 1.1 using the example of the Swiss Roll, which is a 2D manifold embedded in 3D space. Each point lying on the Swiss Roll can be characterized using a 2D parameter vector. Manifold learning techniques deal with the recovery of this underlying parameterization. A number of methods exist for manifold learning; ISOMAP [86], Laplacian Eigenmaps [5] and Locally Linear Embedding [75] are just a few examples. The underlying idea behind most of these methods is to construct a weighted graph, where each data point is represented by a node and the edge weights represent the similarity between a particular pair of nodes. Such graphs are then processed using different methods to identify the underlying parameterization. The output of such methods is the lower-dimensional representation of the data, which preserves certain characteristics of the data in the high-dimensional space. Often, the objective is to preserve the local neighbourhood structure that is present in high-dimensional space. Such techniques always consider full knowledge of the whole dataset. Some studies have also been conducted on techniques for recovery of the data from under-sampled measurements, considering the knowledge of the underlying manifold [22]. However, in most practical problems, the structure of the

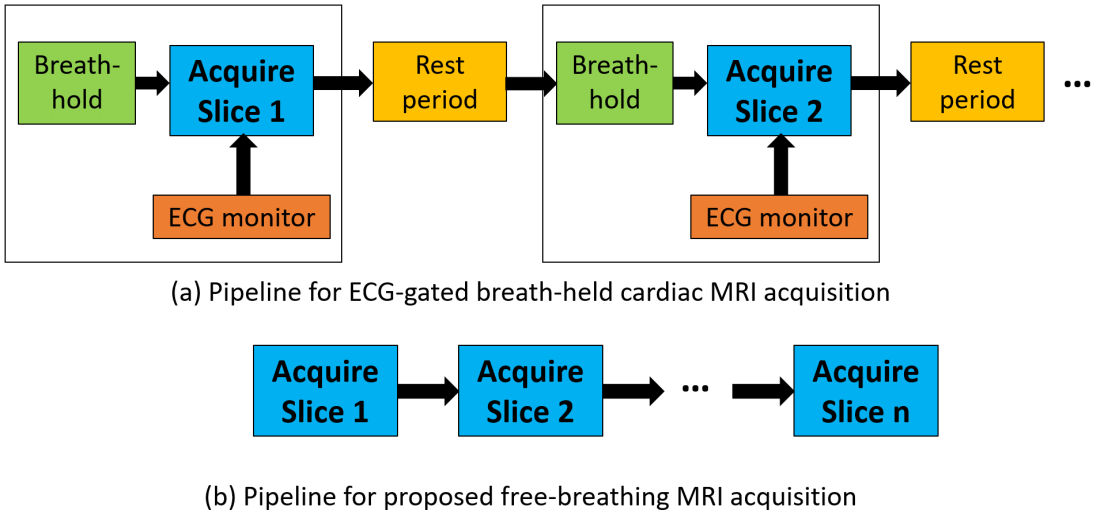


(a) Swiss Roll is a 2D manifold embedded in 3D space

(b) Points on the manifold

Figure 1.1: Points on a low-dimensional manifold: The Swiss Roll shown in (a) is an example of a 2D manifold embedded in 3D space. A number of points are sampled from the Swiss Roll uniformly in (b). Each of these points can be fully characterized by a 2D parameter vector specifying the position of the point on the manifold.

underlying manifold will not be known apriori. Thus, we aim to develop an algorithm that recovers data lying on a manifold from under-sampled measurements, without the prior knowledge of the manifold structure. Our developed algorithm is inspired by the Laplacian Eigenmaps [5] manifold learning algorithm. It aims at recovering the data points by enforcing similarity between neighbouring points in the manifold, much like manifold learning techniques preserve local neighborhoods while computing maps to low-dimensional space. This is achieved by applying fusion penalties to data points within a small neighbourhood on the manifold. We apply the developed algorithm for the recovery of under-sampled data lying on a manifold, to the problem of dynamic cardiac MR image acquisition and reconstruction.



(a) Pipeline for ECG-gated breath-held cardiac MRI acquisition

(b) Pipeline for proposed free-breathing MRI acquisition

Figure 1.2: Acquisition pipelines of gated breath-held and ungated free-breathing cardiac MRI: (a) shows the case where each slice requires a separate breath-hold and the acquisition is synced with the ECG signal. The patient is allowed to rest between breath-holds. In our proposed acquisition scheme in (b), no ECG monitors or breath-holds are required.

1.1.3 Dynamic MR image reconstruction

MRI is a slow imaging modality which collects data in the Fourier domain. Cardiac MR imaging is a challenging problem due to the presence of large cardiac and respiratory motion. Clinically diagnosable imaging quality requires a good spatial and temporal resolution. The desired temporal resolution is around 40 ms, which is too short a time interval to acquire an image of the desired spatial resolution. The usual practice in clinical cardiac cine MRI is to ask the patient to hold his/her breath and then acquire and combine data from multiple heartbeats using ECG gating. The acquisition pipeline is shown in Fig 1.2. Each slice to be acquired requires a separate breath-hold of around 20 s followed by a rest period, which is quite demanding for critically ill patients and paediatric patients. Thus, we propose a free-breathing acquisition and reconstruction scheme which enhances patient comfort and enables the scanning of critically ill patients. This scheme, termed SToRM (Smoothness reg-

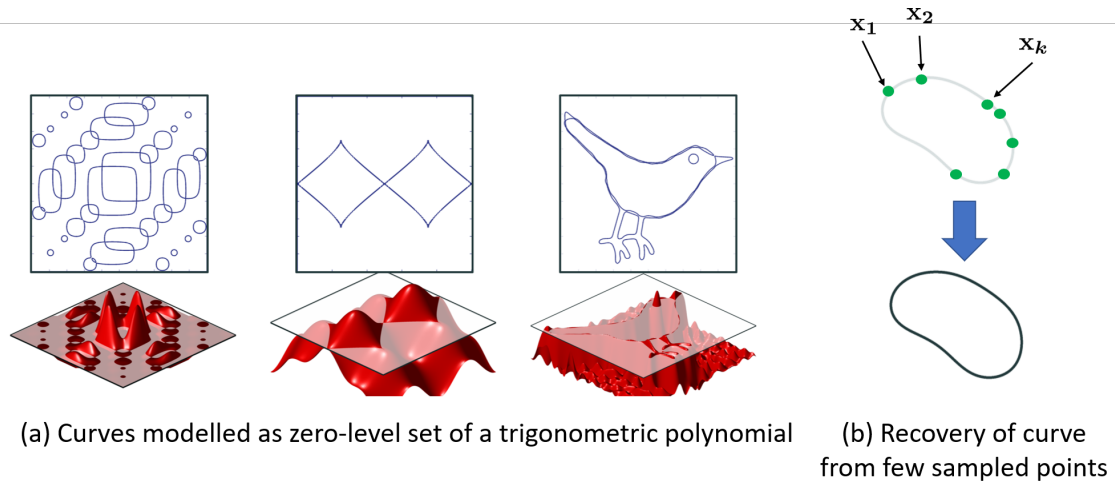


Figure 1.3: Zero-level sets of trigonometric polynomials: (a) shows curves of arbitrary complexity generated as zero-level sets of trigonometric polynomials. (b) shows the problem of recovering the curve uniquely from a few sampled points.

ularization on manifolds) was published in [68]. It shows performance comparable to ECG-gated breath-held techniques, without requiring any physiological monitors or breath-holds. We have also demonstrated that the technique is fairly general, and also showed good performance on the problem of accelerating speech MRI.

1.1.4 Recovery of points on a curve/surface

We next look at the problem of recovery of curves from a few sampled points. We model the curves as the zero-level set of a trigonometric polynomial. This model can represent curves of arbitrary complexity, determined by the bandwidth of the trigonometric polynomial, as shown in Fig 1.3. A practical application which motivated this problem is the reconstruction of DNA filaments from a few clicked points on noisy cryo-electron microscopy images. We show that the matrix of high dimensional feature of points on the curve is rank-deficient, and present techniques to recover the curve from the null-space vectors of this feature matrix. We also derive the sampling conditions required to guarantee unique recovery. The number of samples required

is shown to depend on the bandwidth of the underlying trigonometric polynomial. We demonstrate that the technique is able to recover the DNA filaments from a few points. We next extend the analysis to higher dimensions, where it is computationally infeasible to explicitly form the large feature matrix and find its null-space. In this case, we study the problem of recovery of points satisfying this model from noisy or under-sampled measurements. We formulate the recovery as an optimization problem, where the nuclear norm of the feature matrix is acts as a regularizer. We solve a relaxation of this problem using iterative reweighted algorithms, which only requires the computation of the Gram matrix of the feature matrix. Since the size of the Gram matrix is independent of the ambient dimension of the data, the algorithm is computationally efficient. The algorithm to solve the relaxed optimization problem iterates between the computation of a Laplacian-like matrix and an optimization problem involving fusion penalties. Our proposed model forms the basis for popular kernel low-rank schemes [79]. We show that the proposed scheme is a generalization of the SToRM scheme [68], and show show improved computational efficiency and performance on the cardiac MRI reconstruction problem. This work resulted in the paper [71] and the manuscript [73].

1.1.5 Data arranged in clusters

The problem of clustering data involves grouping different data points such that points within the same group are more similar to each other than to those in other groups. Fig 1.4 shows an example of data which appears in clusters. A real-world example would be to group together different people with similar movie tastes based on their ratings of a common collection of movies. The clustering problem has received considerable attention, which is evident from the huge amount of literature available on the subject. However, classical clustering algorithms such as k-means [32] and spectral clustering [58] suffer from a number of disadvantages such as requiring the prior knowledge of the number of clusters and the sensitivity to initialization due

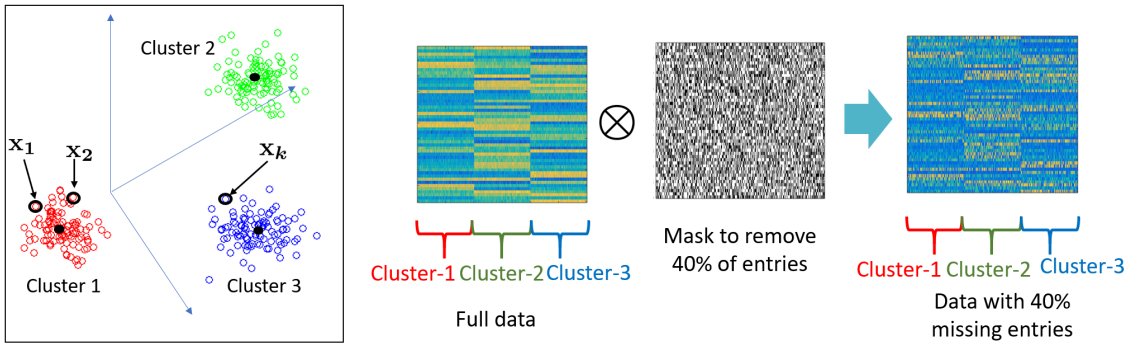


Figure 1.4: Points arranged in clusters: (a) shows an example of 3 distinct clusters in red, green and blue. In (b), a higher dimensional space is considered, and missing feature values are simulated using a mask.

to the non-convexity of the problem being solved. In order to address these issues, recently some convex clustering techniques such as sum-of- norms clustering [34] have been proposed. These algorithms associate an auxiliary variable to each data point, which is to be estimated as the centre of the cluster to which that point belongs. Using fusion penalties as regularizers in the optimization problem, many of these auxiliary variables are estimated to be identical, thus reflecting the cluster structure of the data. However, theoretical guarantees for the various formulations of these convex techniques have not been studied in detail to make inferences about their relative merit. Moreover, these techniques have not been extended to the case where there are missing entries in the available data. We extend the existing sum-of-norms clustering formulation to deal with missing entries, using a $\ell_{2,0}$ norm based fusion penalty. We present theoretical guarantees for correct clustering using the proposed algorithm, and show that the probability of success is higher for well-separated clusters where the cluster membership is not determined by only a few feature values. Since the above problem is NP-hard, we propose a relaxation of the optimization problem using saturating non-convex penalties and present an efficient iterative reweighted

least squares (IRLS) scheme to solve it. The algorithm is demonstrated on simulated as well as real data such as the Wine and the Australian Sign Language (ASL) datasets. It is shown that the proposed scheme is also to detect the clusters present in the datasets even in the presence of a large number of missing entries. This work has resulted in the manuscript [70].

1.2 Contributions

We make several contributions in this thesis to extend traditional compressive sensing algorithms developed for sparse and low-rank models to other data models which are satisfied for real-world data. The algorithms presented are general enough, and we believe that they can also be applied to a variety of other problems which satisfy the same signal models. We list our main contributions below:

1. **Recovery of data on a low-dimensional manifold**: We present an algorithm to recover a series of data, under the assumption that they lie on a low-dimensional manifold embedded in high dimensional space. We devise a novel acquisition scheme which enables us to detect the neighbours of each data point on the manifold. The reconstruction algorithm then proceeds by enforcing similarity between neighbouring points on the manifold. Our technique was inspired by the Laplacian Eigenmaps [5] algorithm, and to the best of our knowledge, we are the first to adapt ideas from such dimensionality reduction algorithms to the reconstruction problem. In addition, we do not require the explicit knowledge of the underlying manifold structure as opposed to other compressive sensing based techniques for this particular data model.

2. **Enabling fast free-breathing dynamic MR scans**: We demonstrate techniques to perform fast dynamic MRI scans with good spatio-temporal resolution in the free-breathing mode, without the need for any physiological monitors. The method is computationally efficient, and has been demonstrated to work

well on a large number of real datasets, and produces images of comparable quality to the clinical standard ECG-gated breath-held scan. This has major clinical significance since many critically ill patients and paediatric patients are unable to perform multiple long breath-holds. We have demonstrated the superiority of our approach over several state-of-the-art techniques. Many of these techniques reconstruct a few cardiac phases as opposed to the full time series recovered by our technique which has more clinical information.

3. **Recovery of curves from few samples**: We formulate the problem as the recovery of the zero-level set of a trigonometric polynomial from a few samples. This model allows the representation of arbitrary, possibly non-smooth curves, whose complexity is determined by the bandwidth of the polynomial. We show that the matrix of high dimensional feature maps of these points is rank deficient, and use the null-space vectors to recover the curves from few measurements. We provide sampling conditions to guarantee perfect recovery of the curves from their samples. The approach is used to recover DNA filaments from a few clicked points on noisy cryo-electron microscopy images.

4. **Denoising and reconstruction of surfaces from few measurements**: We extend the model of zero-level sets of trigonometric polynomials to higher dimensions. We study the problem of denoising or reconstruction from few samples for this data model. We propose to solve an optimization technique which penalizes the nuclear norm of the feature matrix. For such high dimensional data, it is not possible to explicitly form the feature matrix. We show that the problem can be solved efficiently using the Gram matrix of the feature matrix, which is much smaller in size, without having to explicitly form the feature matrix. Thus, our proposed model provides a basis for kernel low-rank based algorithms used in literature.

5. **Clustering of data with missing entries**: We propose a technique to cluster data when a few feature values may be unknown for each data point. Our technique is inspired by the sum-of-norms clustering algorithm [34]. We extend the idea to account for missing entries and provide theoretical guarantees for its success. Moreover, we propose a relaxation of the scheme which is more computationally efficient and use it to cluster real world datasets. The proposed scheme shows good performance even in the presence of a large fraction of missing entries.

1.3 Organization

Chapter 2 introduces a technique for recovery of signals lying on a low-dimensional manifold from a few measurements. This is demonstrated on the problem of free-breathing dynamic MRI acquisition and reconstruction from few Fourier samples. Chapter 3 presents a technique for recovering curves /surfaces from few measurements. The proposed algorithm is used to recover DNA filaments from noisy cryo-electron microscopy data. A computationally efficient algorithm is presented for higher dimensional data. This algorithm is demonstrated on cardiac MRI reconstruction in Chapter 4. Chapter 5 presents a technique for clustering data when a few feature values are unknown for each data point. The technique is analyzed theoretically and a computationally efficient relaxed algorithm is also described. This algorithm is demonstrated on Wine and ASL datasets. Finally, conclusions and future directions are discussed in Chapter 6.

CHAPTER 2
SIGNAL RECOVERY USING MANIFOLD SMOOTHNESS FUSION PENALTIES

2.1 Introduction

We study the problem of recovery of points from their under-sampled measurements, under the assumption that they lie on a low-dimensional manifold. Our idea is to recover the signals by enforcing similarity between the signals in a local neighbourhood on the manifold. Our proposed approach is inspired by the manifold regularization schemes that are widely used in machine learning applications [4, 85, 89]. While our reconstruction algorithm is quite general, we mainly focus on the application of dynamic MR image reconstruction from few Fourier samples. For this particular application, we propose a novel acquisition scheme, which enables us to detect local neighbourhoods on the manifold. The ideas behind this acquisition scheme can be extended to enable the use of our technique on other inverse problems where the same model is satisfied.

Dynamic MR imaging plays a central role in several applications such as structural and functional imaging of the heart, lung and liver, as well as vocal tract imaging in speech. While breath-held and ECG gated imaging is the default acquisition strategy in cardiac MRI, free-breathing un-gated acquisitions can enable the imaging of patients that have difficulty holding their breath [28] (e.g. COPD, obese, and paediatric subjects). Such free running sequences, where the acquisitions are not triggered by physiological signals, can also offer higher acquisition efficiency. The main challenge with free-breathing and ungated strategies (often termed as real-time (RT) imaging), is the slow nature of MR acquisition, which severely restricts the achievable spatial and temporal resolution.

Several model-based reconstruction algorithms that recover dynamic data from undersampled measurements have been introduced to improve the spatial and temporal resolution. The popular approaches include k-t SPARSE methods [39, 51], total

variation (TV) regularization [42], and low rank methods such as k-t PCA [67] or partially separable functions (PSF) [17, 94]. k-t SPARSE methods model the intensity profiles as a sparse linear combination of exponentials. Temporal TV regularization relies on the similarity of each frame with its neighbours in time. PSF and k-t PCA methods exploit the linear dependencies between the intensity profiles by modelling them as a linear combination of basis functions, which are estimated from navigator signals. The main drawback of these schemes in the context of real-time MRI is the degradation in performance with extensive inter-frame motion.

Our proposed scheme, termed SToRM (SmooThness Regularization on Manifolds), exploits the non-linear and non-local dependencies between images in the time series to enable image reconstruction from highly under-sampled measurements. In many RT applications, each image frame in the dataset is a non-linear function of a few physiological parameters (e.g. cardiac and respiratory phase in real-time cardiac cine). Thus the image frames can be modelled as points on a smooth and low dimensional non-linear manifold. Unlike motion resolved reconstruction strategies that bin the data to a few cardiac and respiratory phases and recover them, we propose to recover the entire dynamic dataset from the undersampled k-t data as a manifold smoothness regularized reconstruction problem. We introduce a navigator acquisition scheme to estimate the graph Laplacian matrix. We consider both ℓ_2 and ℓ_1 regularization penalties. We show that the ℓ_2-SToRM formulation can be solved analytically in the Fourier domain in the single receiver coil setting, while it can be solved efficiently using a simple conjugate gradients algorithm in the multi-channel case. We introduce a variable splitting based algorithm to solve for the ℓ_1-SToRM formulation. We demonstrate the utility of our method in accelerated cardiac and speech imaging. The comparisons of the proposed method with the state of the art methods show improved image quality. We expect that our proposed scheme can also be used to accelerate other MR imaging applications such as lung, bowel and liver imaging. The

early version of the proposed algorithm was presented in the conference paper [72], and later extended in [69].

2.2 Background

2.2.1 Manifold regularization

We assume that the points lie on a smooth low-dimensional image manifold i.e. $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k \in \mathcal{M} \subseteq \mathbb{R}^N$. Here $\mathcal{M}$ is a smooth m-dimensional manifold ($m << N$) and N specifies the dimensionality of the signal. The regularized recovery of continuous multi-dimensional functions of a manifold has received considerable attention in the context of machine learning [4]. We present some background on the Laplacian Eigenmaps [5] algorithm to solve this problem, since our work is motivated by this technique. The problem is formulated as:

$$\hat{f} = \arg\min_{f} \mathcal{V}(f) + \lambda \int_{\mathcal{M}} \|\nabla_{\mathcal{M}} f\|^2 d\mathbf{x} \tag{2.1}$$

where f is the continuous function, $\mathcal{V}$ is the desired loss function and $\nabla_{\mathcal{M}} f$ is the derivative of f on $\mathcal{M}$. The second term contains the roughness prior on the manifold which can also be expressed as:

$$\begin{aligned} \int_{\mathcal{M}} \|\nabla_{\mathcal{M}} f\|^2 d\mathbf{x} = \langle \nabla_{\mathcal{M}} f, \nabla_{\mathcal{M}} f \rangle &= \langle f, \triangle_{\mathcal{M}} f \rangle \\ &= \int_{\mathcal{M}} f \ \triangle_{\mathcal{M}} f \ d\mathbf{x} \end{aligned} \tag{2.2}$$

where $\triangle_{\mathcal{M}}$ is the Laplace-Beltrami operator on the manifold. When one is only interested in recovering discrete function values specified by $\mathbf{f} = f_1, f_2, \ldots, f_k$ at points $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k$, the common practice is to approximate the problem as [4]:

$$\hat{\mathbf{f}} = \arg\min_{\mathbf{f}} \mathcal{V}(\mathbf{f}) + \lambda \sum_i \sum_j w_{ij} \ \|f_i - f_j\|^2 \tag{2.3}$$

where the weights w_{ij} are specified by:

$$w_{ij} = \mathbf{e}^{-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{\sigma^2}} \tag{2.4}$$

Note that the weights decay with distance. Specifically, w_{ij} will assume a high value if $\mathbf{x}_i$ and $\mathbf{x}_j$ are similar to each other, and a small value if they are different. The penalty term can also be expressed as:

$$\sum_i \sum_j w_{ij} \|f_i - f_j\|^2 = 2\mathrm{Tr}\left(\mathbf{f}\mathbf{L}\mathbf{f}^H\right) \tag{2.5}$$

where Tr denotes the trace operator and $\mathbf{L}$ is the graph Laplacian operator. The $\mathbf{L}$ matrix is related to the weight matrix $\mathbf{W}$ (with entries defined by (2.4)) as:

$$\mathbf{L} = \mathbf{D} - \mathbf{W} \tag{2.6}$$

where $\mathbf{D}$ is a diagonal matrix with entries $\mathbf{D}(i,i) = \sum_j w_{ij}$. For example, in a three node graph, the Laplacian is:

$$\mathbf{L} = \begin{bmatrix} w_{12} + w_{13} & -w_{12} & -w_{13} \\ -w_{12} & w_{12} + w_{23} & -w_{23} \\ -w_{13} & -w_{23} & w_{13} + w_{23} \end{bmatrix} \tag{2.7}$$

Note the similarity between the discrete approximation (2.5) and (2.2). When the manifold is uniformly sampled, the discrete graph Laplacian operator converges to the Laplace Beltrami operator on the manifold in the limit (as the distance between samples tend to zero) [83]. When $\mathcal{M} = \mathbb{R}^m$, then $\mathbf{L}$ is exactly the finite difference discretization of the continuous Laplacian operator on a regular lattice (up to a constant

factor) [85]:

$$\triangle_{\mathcal{M}} f(\mathbf{r}) = \sum_{i=1}^{m} \frac{f(\mathbf{r}+\mathbf{e}_i) + f(\mathbf{r}-\mathbf{e}_i) - 2f(\mathbf{r})}{\delta^2} = -\frac{[\mathbf{Lf}]_{(\mathbf{r})}}{\delta^2} \tag{2.8}$$

where $\mathbf{e}_1, \ldots \mathbf{e}_m$ form an orthogonal basis for $\mathbb{R}^m$ with $\|\mathbf{e}_i\| = \delta$.

2.2.2 Acquisition scheme in MRI

We now briefly summarize the MR image acquisition process. We model the raw dynamic multi-channel MRI data from the i^{th} image frame $\mathbf{x}_i$ as:

$$\mathbf{b}_{ij} = \underbrace{\mathbf{S}_i \mathbf{F} \mathbf{C}_j}_{\mathbf{A}_{ij}} \mathbf{x}_i + \boldsymbol{\eta}_{ij}, \quad j = 1, .., \mathrm{N}_{\text{coils}} \tag{2.9}$$

where $\mathbf{C}_j$ is the receive sensitivity of the j^{th} coil, $\mathbf{S}_i$ is the sampling pattern for the i^{th} frame and $\boldsymbol{\eta}_{ij}$ is the noise. $\mathbf{F}$ is the discrete Fourier transform matrix. For our problem of dynamic MR image reconstruction, the operator $\mathbf{S}_i$ selects very few samples from the Fourier transform of each frame, resulting in highly under-sampled measurements. The above can be simplified and re-written as:

$$\mathbf{B} = \mathcal{A}(\mathbf{X}) + \boldsymbol{\eta} \tag{2.10}$$

where $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_k]$ is the Casorati matrix obtained by stacking the vectorized images as columns.

2.3 Proposed scheme

We model the signals in the dataset as points on a smooth low-dimensional manifold parameterized by a few variables. For example, the images in a free-breathing and ungated cardiac MRI dataset are non-linear functions of their cardiac and respiratory phases. The proposed framework is general enough to be applied to several dynamic imaging applications like imaging of the vocal tract in speech, where there

is no concept of phases equivalent to cardiac and respiratory phases in cardiac imaging. We propose to recover the signals from their undersampled measurements (2.9) by exploiting the manifold structure of the data. Motivated by (2.3), we pose the recovery as:

$$\{\mathbf{X}^*\} = \arg\min_{\mathbf{X}} \|\mathcal{A}(\mathbf{X}) - \mathbf{B}\|_F^2 + \lambda \sum_i \sum_j \left(\sqrt{w_{ij}}\ \|\mathbf{x}_i - \mathbf{x}_j\|_p\right)^p \tag{2.11}$$

where we use the ℓ_p ($p \leq 2$) norm of the signal differences in the regularizer. We will consider the special cases $p = 2$ and $p = 1$ in the later subsections. The above optimization problem promotes solutions where each data point is similar in the lp norm sense to its neighbours on the manifold and the degree of similarity is determined by the weights w_{ij}.

In classical manifold embedding applications, the weights are derived from the signals themselves. This approach is not practical in our setting since we only have a few measurements available from each signal. We now present a technique to estimate these weights for the case of dynamic MR imaging. The idea can be extended to other applications where there is some flexibility in controlling the acquisition procedure. Our acquisition strategy is similar to [17, 94], and uses navigators to estimate the weights to be used in (2.11).

2.3.1 Estimation of manifold structure from navigators

Consider that each of the k images is observed by the same $M \times N$ matrix $\mathbf{\Psi}$ ($M < N$). This mapping is a stable embedding if the distance between any two points $\mathbf{x}_i$ and $\mathbf{x}_j$ is preserved after the mapping $\mathbf{\Psi}\mathbf{x}_i$. Wakin et al [22] have shown that a random orthoprojector $\mathbf{\Psi}$ provides a stable embedding of the manifold. Specifically, for some $0 < \epsilon \leq \frac{1}{3}$ and a sufficient number of measurements M, the following holds

with high probability for every pair $\mathbf{x}_i$ and $\mathbf{x}_j$:

$$(1-\epsilon)\|\mathbf{x}_i-\mathbf{x}_j\| \leq \|\mathbf{\Psi}\mathbf{x}_i-\mathbf{\Psi}\mathbf{x}_j\| \leq (1+\epsilon)\|\mathbf{x}_i-\mathbf{x}_j\| \tag{2.12}$$

The number of measurements M required to yield stable embedding is independent of the ambient dimension N and is almost linearly proportional to the dimension of the manifold m. The required number of measurements also depends on the characteristics of the manifold which are captured by its condition number and volume [59].

Motivated by the above result, we propose to sample the same k-space locations in every temporal frame. We term the common measurements as navigator acquisitions, which are often used in many dynamic MRI applications for calibration [17, 94]. We define the measurement operator $\mathbf{A}_{ij}$ corresponding to the i^{th} frame and the j^{th} coil as (see (2.9)):

$$\mathbf{b}_{i,j} = \underbrace{\begin{bmatrix} \mathbf{\Phi} \\ \\ \mathbf{B}_i \end{bmatrix} \mathbf{F}\ \mathbf{C}_j}_{\mathbf{A}_{ij}}\ \mathbf{x}_i + \boldsymbol{\eta}_{ij} \tag{2.13}$$

The first operator $\mathbf{\Phi}$ samples the same k-space locations every frame, regardless of the frame number i; the corresponding samples (termed navigator signals) enable the estimation of the neighbours of each frame. The second operator $\mathbf{B}_i$ which samples different k-space locations every frame aids the image recovery algorithm by sampling the neighbours of a particular image frame at complementary k-space locations. We propose to estimate the inter-image distances as:

$$d_{ij}^2 = \sum_{l=1}^{\mathrm{N_{coils}}} \|\mathbf{z}_{il}-\mathbf{z}_{jl}\|^2 \tag{2.14}$$

where $\mathbf{z}_{i,l}$ are the navigator signals given by:

$$\mathbf{z}_{i,l} = \mathbf{\Phi}\mathbf{F}\ \mathbf{C}_l\ \mathbf{x}_i + \boldsymbol{\eta}_{il},\ l = 1,..,\mathrm{N_{coils}} \tag{2.15}$$

We compute the weights as:

$$w_{ij} = \begin{cases} \mathbf{e}^{-\frac{d_{ij}^2}{\sigma^2}} & \text{, if } \mathbf{x}_i \text{ and } \mathbf{x}_j \text{ are neighbours.} \\ 0 & \text{, otherwise.} \end{cases} \tag{2.16}$$

We set the neighbourhood of each frame to be a fixed number of nearest neighbours. For example, in order to retain the 5 nearest neighbours for each frame, the i^{th} and the j^{th} frames are considered to be neighbours if the i^{th} frame is among the 5 frames most similar to the j^{th} frame or the j^{th} frame is among the 5 frames most similar to the i^{th} frame.

2.3.2 Special case: ℓ_2 smoothness prior

When $p = 2$, the recovery using (2.11) simplifies to:

$$\mathbf{X}^* = \arg\min_{\mathbf{X}} \|\mathcal{A}(\mathbf{X}) - \mathbf{B}\|_F^2 + 2\lambda \ \mathrm{Tr}(\mathbf{X}\mathbf{L}\mathbf{X}^H), \tag{2.17}$$

where the Laplacian matrix $\mathbf{L}$ is obtained from the weights using (2.6). We refer to this implementation as ℓ_2-SToRM. We can view (2.17) as an analysis formulation since the regularizer is based on the analysis of $\mathbf{X}$ (specified by $\mathbf{X}\mathbf{Q}$), where $\mathbf{L} = \mathbf{Q}\mathbf{Q}^H$. The problem (2.17) can be rewritten as:

$$\{\mathbf{X}^*\} = \arg\min_{\mathbf{X}} \|\mathcal{A}(\mathbf{X}) - \mathbf{B}\|_F^2 + 2 \ \lambda \|\mathbf{X}\mathbf{Q}\|^2. \tag{2.18}$$

The $k \times k(k-1)/2$ matrix $\mathbf{Q}$ specifies a gradient operator. For example, in a 4 node graph, the matrix $\mathbf{Q}$ is specified by:

$$\mathbf{Q}^T = \begin{bmatrix} \sqrt{w}_{12} & -\sqrt{w}_{12} & 0 & 0 \\ 0 & \sqrt{w}_{23} & -\sqrt{w}_{23} & 0 \\ 0 & 0 & \sqrt{w}_{34} & -\sqrt{w}_{34} \\ \sqrt{w}_{13} & 0 & -\sqrt{w}_{13} & 0 \\ 0 & \sqrt{w}_{24} & 0 & -\sqrt{w}_{24} \\ \sqrt{w}_{14} & 0 & 0 & -\sqrt{w}_{14} \end{bmatrix} \tag{2.19}$$

Note that this approach is very similar to Tikhonov temporal regularization, when the sparse matrix $\mathbf{Q}$ is the temporal finite difference operator. The proposed scheme uses an operator that computes differences between the neighbours on the manifold, rather than the temporal neighbours. Since the neighbours on the manifold are expected to be more similar than the ones in time, we expect to obtain better recovery.

We will now show that this formulation is also equivalent to a synthesis formulation by a simple change of variables. In addition to providing additional insights, this offers an approach to represent the data efficiently, while working with large datasets. The Laplacian matrix has a singular value decomposition specified by:

$$\mathbf{L} = \mathbf{V}\mathbf{\Sigma}\mathbf{V}^H \tag{2.20}$$

The eigen vectors of the Laplacian matrix denoted by $\mathbf{v}_i$ are ideally suited to represent smooth signals on the manifold. A simple special case worth discussing is when the graph has r disjoint clusters. In this case, spectral graph theory shows that $\mathbf{L}$ will have r zero singular values. The corresponding r singular vectors $\mathbf{V}_0$ with an appropriate

rotation matrix $\mathbf{R}$ will yield a set of sparse temporal basis functions:

$$\mathbf{E}_0 = \mathbf{R}\mathbf{V}_0 \tag{2.21}$$

Each of the basis functions in $\mathbf{E}_0$ will assume a value of zero for frames that are not in a particular cluster, and a constant value for all the frames in the cluster. This property is exploited in spectral clustering. If the images in the cluster are the same, these temporal basis functions are sufficient to represent the signal. Note that this representation is strikingly different from principle component analysis used in k-t PCA or PSF methods [17, 94]. Unlike the global subspace model used in these methods, the proposed approach captures the geometry of the data on the manifold, enabled by the non-linear mapping (2.16). By minimizing the cross talk between images in distinct cardiac/respiratory phases, it is expected to reduce temporal blurring.

In the general setting, one would need more basis functions to account for the variability of images within clusters/on the manifold. Substituting $\mathbf{L}$ in the regularization penalty term in (2.17), we obtain:

$$\mathrm{Tr}\left(\underbrace{(\mathbf{X}\mathbf{V})}_{\mathbf{U}}\mathbf{\Sigma}(\mathbf{X}\mathbf{V})^H\right) = \sum_{i=1}^{k}\sigma_i\|\mathbf{u}_i\|^2, \tag{2.22}$$

where $\mathbf{u}_i = \mathbf{X}\mathbf{v}_i$ is the projection of $\mathbf{X}$ onto the i^{th} singular vector $\mathbf{v}_i$ and σ_i is the i^{th} singular value of $\mathbf{L}$. Substituting for $\mathbf{X}$ in terms of $\mathbf{U}$ in (2.17), we obtain the equivalent synthesis formulation:

$$\mathbf{U}^* = \arg\min_{\mathbf{U}} \|\mathcal{A}(\mathbf{U}\mathbf{V}^H) - \mathbf{B}\|_F^2 + 2\lambda \sum_{i=1}^{k}\sigma_i\|\mathbf{u}_i\|^2 \tag{2.23}$$

Note that the above formulation is very similar to the k-t PCA or PSF [17] algorithms that are now widely used in dynamic MRI. The columns of $\mathbf{U}$ correspond to representative images, while the columns of $\mathbf{V}$ are the corresponding temporal basis

functions.

2.3.3 Special case: ℓ_1 smoothness prior

We consider the ℓ_1 norm of the differences between neighbouring images on the manifold:

$$\{\mathbf{X}^*\} = \arg\min_{\mathbf{X}} \|\mathcal{A}(\mathbf{X}) - \mathbf{B}\|_F^2 \ + 2\lambda \|\mathbf{X}\mathbf{Q}\|_1 \tag{2.24}$$

We term this implementation ℓ_1-SToRM. Note that the above approach simplifies to the popular temporal total variation formulation when:

$$w_{ij} = \begin{cases} 1, & \text{if } \ j = i+1, i-1. \\ 0, & \text{otherwise.} \end{cases} \tag{2.25}$$

We expect our method to achieve better reconstruction than temporal TV since it enforces the differences between the closest neighbours of a frame on the manifold to be sparse. These frames might not be the frames that are close to it temporally, especially in case of high motion between frames.

Considering that $\mathbf{Q}$ has a singular value decomposition:

$$\mathbf{Q} = \mathbf{V}\mathbf{\Sigma}^{\frac{1}{2}}\mathbf{T} \tag{2.26}$$

we can also find the equivalent synthesis formulation for the ℓ_1 problem by a change of variable $\mathbf{X} = \mathbf{U}\mathbf{V}^H$:

$$\mathbf{U}^* = \arg\min_{\mathbf{U}} \|\mathcal{A}(\mathbf{U}\mathbf{V}^H) - \mathbf{B}\|_F^2 + 2\lambda \ \|\mathbf{U}\mathbf{\Sigma}^{\frac{1}{2}}\mathbf{T}\|_1, \tag{2.27}$$

Note that this approach has similarities to ℓ_1 regularized PSF regularization schemes [94], except that the ℓ_1 norm of $\mathbf{U}\mathbf{\Sigma}^{\frac{1}{2}}\mathbf{T}$ is penalized rather than that of $\mathbf{U}$.

2.4 Implementation

We consider separately the solutions for the 2 cases described in the previous section: $p = 2$ and $p = 1$.

2.4.1 ℓ_2 smoothness prior

In the single coil case, problem (2.17) has an analytical solution in the Fourier domain. We rewrite (2.17) in this special case as:

$$\hat{\mathbf{X}}^* = \arg\min_{\hat{\mathbf{X}}} \sum_i \|\mathbf{S}_i \hat{\mathbf{x}}_i - \mathbf{b}_i\|_F^2 + 2\lambda \mathrm{Tr}(\, \hat{\mathbf{X}}\, \mathbf{L}\, \hat{\mathbf{X}}^H) \tag{2.28}$$

where the columns of $\hat{\mathbf{X}} = [\hat{\mathbf{x}}_1, \ldots, \hat{\mathbf{x}}_k]$ are the Fourier coefficients of the images given by: $\hat{\mathbf{x}}_i = \mathbf{F}\mathbf{x}_i$. The key observation is that the above expression can be decoupled into several independent subproblems, each involving the recovery of a row of $\hat{\mathbf{X}}$. Let $\hat{\mathbf{X}}_{(j,\cdot)}$ denote the j^{th} row of $\hat{\mathbf{X}}$ and $\mathbf{B}_{(j,\cdot)}$ denote the vector of measurements corresponding to this row. Then, we can solve for $\hat{\mathbf{X}}_{(j,\cdot)}$ analytically as:

$$\hat{\mathbf{X}}_{(j,\cdot)} = (\mathbf{D}_j^H \mathbf{D}_j + 2\lambda \mathbf{L})^{-1} \mathbf{D}_j^H\, \mathbf{B}_{(j,\cdot)} \tag{2.29}$$

where $\mathbf{D}_j$ is the sampling matrix corresponding to the j^{th} row. The solutions for the different rows of $\hat{\mathbf{X}}$ can be computed in parallel. This analytic approach can give us a significant speed-up over solving for the whole matrix $\hat{\mathbf{X}}$ using iterative algorithms such as conjugate gradient.

In the multi-channel setting, it is possible to solve for each coil using the above method and combine them using a sum-of-squares strategy. Since this approach is suboptimal, we propose to directly solve (2.17) using the conjugate gradient algorithm (accounting for the coil sensitivities) to obtain a more accurate solution. The gradient of the cost function in (2.17) can be computed as: $2\mathcal{A}^H\mathcal{A}(\mathbf{X}) + 4\mathbf{X}\mathbf{L}$. The computation of $\mathcal{A}^H\mathcal{A}(\mathbf{X})$ can be broken down into blocks (each containing a few temporal frames

of $\mathbf{X}$) and the blocks can be processed in parallel in order to reduce computational complexity.

2.4.2 ℓ_1 smoothness formulation

We rely on a variable splitting strategy using an auxiliary variable $\mathbf{Z}$ to solve (2.24):

$$\begin{aligned}\{\mathbf{X}^*, \mathbf{Z}^*\} &= \arg\min_{\mathbf{X},\mathbf{Z}} \|\mathcal{A}(\mathbf{X}) - \mathbf{B}\|_F^2 + \\ &\qquad 2\lambda\|\mathbf{Z}\|_{\ell_1} + \beta\|\mathbf{XQ} - \mathbf{Z}\|_F^2 \end{aligned} \tag{2.30}$$

We solve the above problem by alternating between minimization with respect to the 2 variables:

$$\begin{aligned}\mathbf{X}^{(n)} &= \arg\min_{\mathbf{X}} \ \|\mathcal{A}(\mathbf{X}) - \mathbf{B}\|_F^2 + \\ &\qquad \beta\|\mathbf{XQ} - \mathbf{Z}^{(n-1)}\|_F^2 \end{aligned} \tag{2.31}$$

$$\mathbf{Z}^{(n)} = \arg\min_{\mathbf{Z}} \ \beta\|\mathbf{X}^{(n)}\mathbf{Q} - \mathbf{Z}\|_F^2 + 2\lambda\|\mathbf{Z}\|_1 \tag{2.32}$$

We use a homotopy continuation strategy on the parameter β, where β is initialized to a very small value and then increased gradually to a very large value till the algorithm converges. As in the ℓ_2 case, (2.31) can be solved analytically in the Fourier domain for single coil data. For multi-coil data, we use the conjugate gradient algorithm. (2.32) can be solved using shrinkage. The matrix $\mathbf{Z}$ is large and storing it explicitly will result in huge memory demands. We observe that the evaluation of (2.31) only requires $\mathbf{ZQ}^T$, which is considerably smaller in dimension than $\mathbf{Z}$. We perform in-place computation of the variable $\mathbf{ZQ}^T$ and store it instead of $\mathbf{Z}$ to reduce the memory demand of the algorithm.

2.4.3 Acquisition scheme

The acquisition scheme used follows from the discussion in section 2.3.1. We used a set of uniformly spaced radial navigator acquisitions (corresponding to $\mathbf{\Phi}$), separated by $180^\circ/N_l$ degrees where N_l is the number of navigator lines per frame. The remaining k-space samples (corresponding to $\boldsymbol{B}_i$) were acquired using a golden angle radial k-space trajectory, where each line was separated by an angle of 111.25° from the previous line. Thus, $\boldsymbol{B}_i$ varies from frame to frame. The acquisition and reconstruction pipeline is illustrated in Fig 2.1, where we consider the single coil setup for simplicity.

2.4.4 Datasets

We use a numerical cardiac phantom and a retrospectively undersampled speech dataset for quantitative comparisons. We also consider the recovery of prospectively undersampled real-time cardiac MRI data.

2.4.4.1 PINCAT phantom

A short axis view of the PINCAT phantom [80] heart with matrix size 128×128 and 500 frames was used for numerical simulations. The dataset has around 26 cardiac cycles and 5 respiration cycles.

2.4.4.2 Speech imaging

We use the MR dataset titled 'F1' in the USC-TIMIT database [57] to demonstrate our method. The raw k-space data for the images in the database was acquired using a spiral trajectory and this data was gridded to reconstruct the images. The reconstructed images have been made available in the dataset as a movie in the coil-combined form with matrix size 68×68 and frame-rate 23.18 frames/s. This corresponds to a temporal resolution of around 43 ms. The Fourier data corresponding to the first 6000 image frames was retrospectively undersampled using 9 golden angle

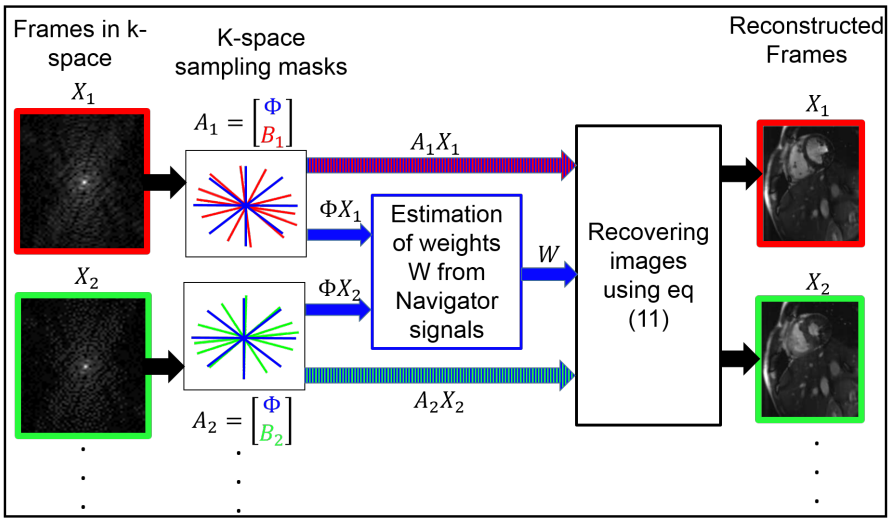


Figure 2.1: Summary of the proposed data acquisition and reconstruction scheme for the single coil case. The blue radial lines denote the navigators that sample the same k-space locations in every frame. The weight matrix is estimated from the k-space data acquired using these navigator lines as described in (2.16). The final images are recovered from the entire measurements by solving (2.11).

radial lines and 1 spiral navigator per frame and used for our experiments.

2.4.4.3 Cardiac Imaging

A prospectively undersampled free-breathing ungated radial dataset was acquired using a SSFP sequence on a Siemens 3T TIM Trio scanner with a 18 channel cardiac array from a healthy volunteer who was asked to breathe normally. The scan parameters were TR/TE = 4.2/2.2 ms, number of slices = 5, slice thickness = 5 mm, FOV = 300 mm, spatial resolution = 1.17 mm. A temporal resolution of 42 ms was achieved by sampling 10 lines of k-space per frame, out of which 4 were navigator lines. 10000 radial lines of k-space were acquired per slice which resulted in an acquisition time of around 42 s per slice. For 5 slices this resulted in a total acquisition time of around 3.5 mins.

The raw k-space data was interpolated to a Cartesian grid and a SVD based coil-compression technique was used in order to create 4 virtual coil elements from the

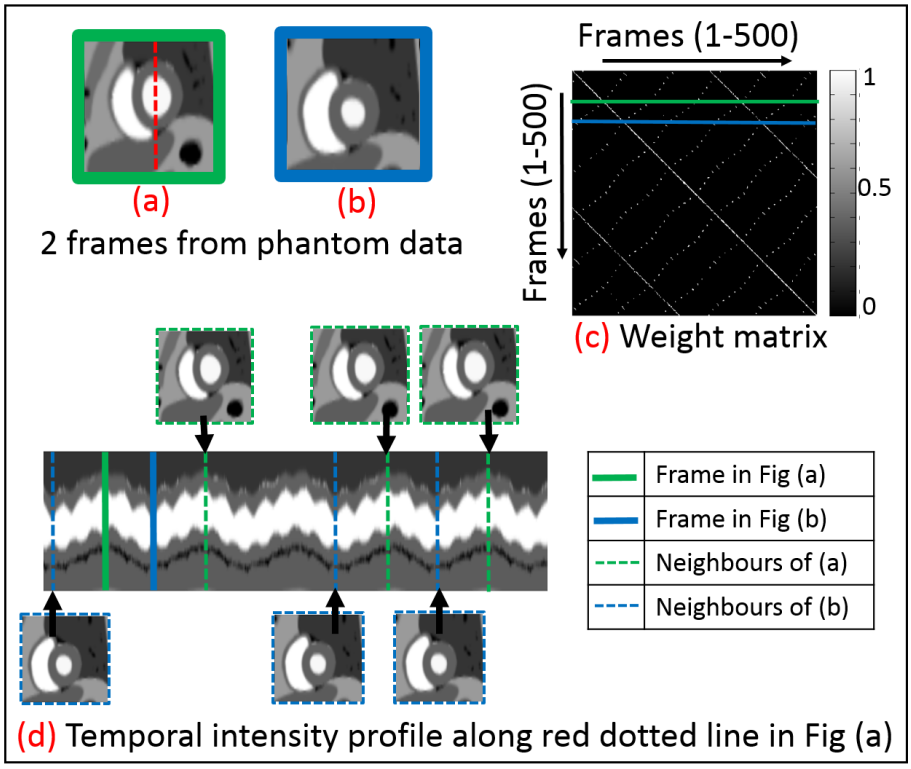


Figure 2.2: Illustration of the weight matrix and the ability of the scheme to enable implicit motion resolved recovery. **(a,b)** Two frames from the PINCAT dataset. **(c)** Weight matrix computed from the fully sampled k-space data. The green and blue lines show the rows corresponding to the frames in (a) and (b) respectively. The neighbours of these frames can be obtained using the weight matrix. **(d)** Temporal intensity profile corresponding to the cut shown by the red dotted line in (a). Frames (a) and (b) and a few of their neighbours are marked.

initial 18. We reconstructed low temporal resolution images for the original coils by binning k-space data from a large number of frames. We then performed an SVD on these images and retained only the 4 most significant singular vectors. The data from the original coils was coil-combined to form virtual coil data using the singular vectors obtained. This was done in order to reduce the computational complexity of the reconstruction procedure. The coil sensitivity maps were estimated from this compressed data using the method by Walsh et al [91]. To reduce computational complexity, the coil sensitivity maps were assumed to be constant over time.

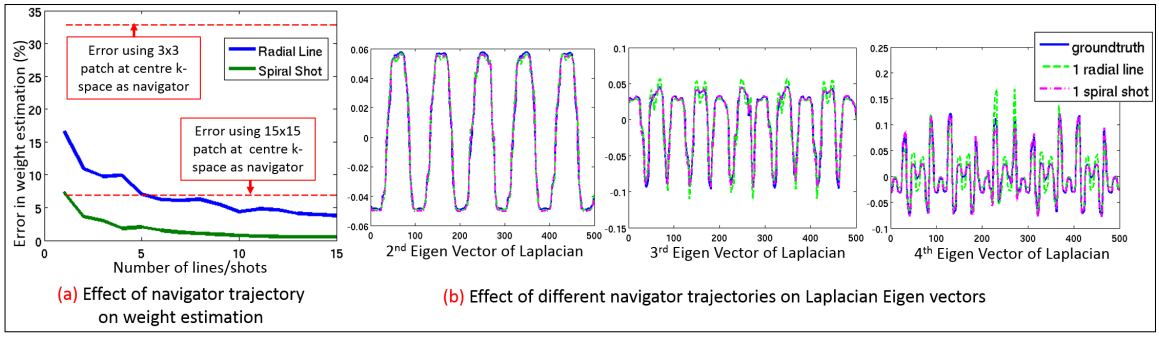


Figure 2.3: Effect of different navigator trajectories on weight matrix estimation. **(a)** Percentage error in the weight matrix estimation (computed using ℓ_2 norm), using different navigator trajectories. Spiral and radial trajectories are chosen such that the time taken to acquire 1 spiral shot is the same as that for 1 radial line. **(b)** The $2^{\text{nd}}, 3^{\text{rd}}$ and 4^{th} eigen vectors of the Laplacian matrix estimated from (1) fully sampled k-space, shown in blue (2) 1 radial spoke, shown in green (3) 1 spiral readout, shown in pink. We observe that these vectors capture the respiratory motion, the 2nd harmonic of the respiratory motion, and the cardiac motion modulated by the respiratory frequency respectively.

2.4.4.4 Comparison between breath-held and free-breathing cardiac acquisitions

In order to compare the image quality obtained using our method to that obtained by a breath-held protocol, we acquired 2 cardiac datasets:

- A prospectively undersampled free-breathing ungated radial dataset.
- A fully-sampled breath-held ECG-gated radial dataset.

The first dataset was acquired using a SSFP sequence on a Siemens 3T TIM Trio scanner with a 5 channel cardiac array from a healthy volunteer who was asked to breathe normally. A TRUFI frequency scout was performed prior to data acquisition to prevent banding artifacts due to the presence of field in-homogeneity. The scan parameters were TR/TE = 3.2/1.62 ms, number of slices = 5, slice thickness = 5 mm, FOV = 300 mm, spatial resolution = 1.17 mm. A temporal resolution of 41.6 ms was achieved by sampling 13 lines of k-space per frame, out of which 4 were navigator lines. 13000 radial lines of k-space were acquired per slice which resulted

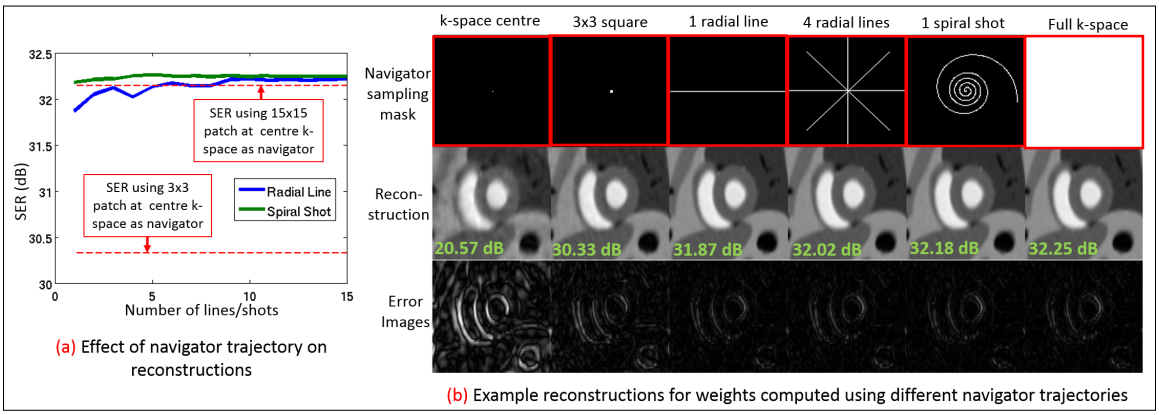


Figure 2.4: Effect of weight matrices estimated using different navigator trajectories on reconstruction. **(a)** Signal to error ratio of the reconstructions with the Laplacian matrix estimated from different navigator trajectories. The k-space samples used to reconstruct the images are the same for all cases (10 golden angle radial lines per frame). Only the navigator trajectory used to compute the weight matrix are varied. **(b)** A reconstructed frame is shown for a few of the trajectories reported in (a).

in an acquisition time of around 42 s per slice. For 5 slices this resulted in a total acquisition time of around 3.3 mins.

The fully-sampled ECG-gated breath-held dataset was acquired by a SSFP sequence on the same subject immediately after the free-breathing scan. The sampling trajectory was uniform radial and the scan parameters were: TR/TE = 3.4/1.72 ms, number of slices = 5, slice thickness = 5 mm, number of channels = 5, FOV = 300 mm, spatial resolution = 1.17 mm, number of cardiac phases = 18, radial views per cardiac phase = 253. Each slice required a breath-hold of around 16 s followed by a resting period of around 25 s. For 5 slices this resulted in a total acquisition time of around 3 mins.

Pre-interpolation to a Cartesian grid, coil sensitivity estimation and coil compression were performed using the acquired k-space data as described in the previous section. 3 virtual coils were created in this case.

2.4.5 State of the art methods used for comparison

The in vivo data reconstructed using ℓ_2 and ℓ_1-SToRM was compared to the reconstructions by 3 other methods: kt-LR [48], temporal TV and PSF. The kt-LR and temporal TV methods do not require the acquisition of navigators. Thus, we did not include navigator lines in our sampling pattern for the speech data, for generating the results for these 2 methods. However, we could not do the same for the cardiac datasets since they were prospectively undersampled. For the PSF method, we used the Frobenius norm of the basis images as a regularizer. The approach followed was similar to [17], with the same weighting applied to all basis images. For all 3 competing methods, the regularization parameter giving the highest SER reconstruction was chosen in case of the speech dataset. For the cardiac dataset, since the ground-truth was not available, the regularization parameter which seemed to best preserve the features of the data was chosen. Spatial TV regularization was not used with any of the algorithms.

2.5 Results

2.5.1 Simulations using phantom data

We first conducted some numerical simulations on the PINCAT phantom. Two frames of the phantom dataset are shown in Fig 2.2.(a) and Fig 2.2.(b).

2.5.1.1 Weight matrix estimate from fully sampled data

We computed the weight matrices from the fully sampled k-space data, corresponding to different σ values. These matrices were thresholded to retain only the 5 nearest neighbours for each frame. The k-space data was then under-sampled (10 lines per frame sampled on a pseudo golden angle trajectory). Images were reconstructed from this under-sampled data using ℓ_2-SToRM with the weight matrices corresponding to different σ values. The σ value giving the highest SER reconstruction was chosen to form the optimal weight matrix. This matrix is shown in Fig 2.2.(c). The

temporal intensity profile of the original dataset (along the cut given by the red dotted line in Fig 2.2.(a)) is shown in Fig 2.2.(d). The frames in Fig 2.2.(a) and Fig 2.2.(b) and a few of their neighbours (obtained from the weight matrix) are marked along the profile. We observe that the frames in Fig 2.2.(a) and Fig 2.2.(b) are very similar to their neighbours estimated by the weight matrix.

2.5.1.2 Effect of navigator trajectory on weight matrix estimation

The effect of different navigator schemes on weight estimation is studied in Fig 2.3. The weights estimated from different trajectories were compared quantitatively to the ground-truth weights obtained from the fully sampled data (Section 2.5.1.1). The normalized ℓ_2 norm of the weight estimation error was used as the error metric. The optimal σ parameter varies from trajectory to trajectory, depending on the number of k-space points. We chose the best σ value in each case to obtain fair comparisons. We did not threshold the weight matrices for this experiment. We considered spiral and radial navigators with the same readout duration (TR = 4.3 ms). The percent errors in weight estimation (computed using ℓ_2 norm) are plotted in Fig 2.3.(a). We observe that 1 spiral shot (4.3 ms) is almost as accurate in estimating the weights as 5 radial lines (21.5 ms). The percent errors incurred in the two cases are 7.34% and 6.99% respectively. The $2^{\text{nd}}, 3^{\text{rd}}$ and 4^{th} eigen-vectors of the $\mathbf{L}$ matrix estimated from: (1) the fully sampled data, (2) 1 radial line, and (3) 1 spiral shot are shown in Fig 2.3.(b). The experiments show that the eigen vectors in all three cases are quite similar. We also observe that the 2^{nd} eigen-vector captures the respiratory motion of the data (5 respiratory cycles can be seen). The 3^{rd} eigen-vector shows the 2^{nd} harmonic of the respiratory motion. The dominant frequency of this eigen-vector is double that of the dominant frequency of the respiratory motion. The 4^{th} eigen vector captures the cardiac motion modulated by the respiratory frequency (26 cardiac cycles can be seen).

2.5.1.3 Effect of weight matrix on image reconstruction

The effect of different weight matrices (computed using the navigator trajectories described in Section 2.5.1.2) on image reconstruction quality is studied in Fig 2.4.(a). The phantom data was under-sampled in k-space using a golden angle trajectory with 10 lines per frame and this data was used for all reconstructions. The navigator data was used only for weight computation. The weight matrices were thresholded to retain only the 5 nearest neighbours for each frame. In Fig 2.4.(b), we show a single image frame from the time series, as reconstructed using different weight matrices. The weights computed using a 1 radial line navigator produced reconstructed images of comparable quality (31.87 dB) to the case of ground-truth weights (32.25 dB). The single shot spiral navigator trajectory, which takes the same acquisition time as 1 radial line, performed slightly better (32.18 dB) than the single radial line case. Estimation of weights using only the centre k-space signal gave very poor reconstructions (20.57 dB). Using a 3×3 patch around center k-space as the navigator signal (instead of the centre only) improved the results considerably (30.33 dB), though the error images show more artifacts than when using radial or spiral trajectories.

We clarify that for the above experiment we used the navigator data only for estimating the weights and not for reconstruction. However, the navigator data was used for reconstruction in all the subsequent in-vivo experiments on the speech and cardiac data. For the experiment in Sec 2.5.1.2, we were studying the relative merits of different sampling schemes on the weight computation. The analysis was extended in the above experiment, where we studied the effect of those computed weights on image reconstruction. If we included the navigator signals for the reconstruction step, then the quality of our reconstructed images would be dependent on: (1) The accuracy of the computed weights (2) The incoherence of the sampling patterns used for each trajectory. Since we were only studying effect (1), we used the same samples for reconstruction in each case.

2.5.2 Experiments on in vivo data

In the in vivo experiments, the parameter σ used for the calculation of the weight matrix $\mathbf{W}$ was automatically computed using the strategy described in [84]. For this purpose, we computed the weight matrix for a range of σ values and evaluated $l(\sigma) = \sum_{i=1}^{k}\sum_{j=1}^{k} W_{ij}(\sigma)$ for each weight matrix obtained. A log-log plot of $l(\sigma)$ revealed 2 constant asymptotes at $\sigma \to 0$ and $\sigma \to \infty$, smoothly connected by an approximately straight line. The approximate σ value at the middle of this linear portion was selected to form our weight matrix. The weight matrix was thresholded to retain only the 4-6 nearest neighbours for each frame. For the multi-slice cardiac datasets, we had to compute the weight matrix separately for each slice. This is because our acquisition scheme was 2D, i.e. the k-space samples from a particular slice were completely acquired before moving on to the next slice. The regularization parameter λ was chosen emperically. All reconstructions were done on a desktop computer (Intel Xeon E5-1620 CPU, 3.6 GHz, 32 GB RAM). The algorithms were implemented in MATLAB, and may be further optimized to produce lower reconstruction times.

2.5.2.1 Retrospective undersampling experiments on speech dataset

The speech dataset was reconstructed from under-sampled k-space data using different techniques, as shown in Fig 2.5. The first row (a) shows the ground-truth images, while each subsequent row corresponds to datasets reconstructed by different methods. The techniques used for reconstruction from under-sampled k-space data along with their reconstruction times are: (b) kt-LR (4.8 hrs) (c) temporal TV (21 mins) (d) PSF [17] (3 mins) (e) ℓ_2-SToRM (7 mins) (f) ℓ_1-SToRM (32 mins). For the kt-LR and temporal TV reconstructions, 10 golden angle radial lines of k-space were used per frame and no navigator lines were included. For the PSF and SToRM methods, k-space was undersampled using 9 golden angle radial lines and 1 spiral navigator per frame. SToRM produces reconstructions with higher SER than the other methods. Though the ℓ_1 and ℓ_2-SToRM reconstructions have comparable SER, it can be

seen than the ℓ_1 formulation reduces blurring and preserves borders better. The competing techniques have more artifacts compared to the proposed methods, as pointed out in the figure. The ability to recover high quality images from under-sampled data indicates that our method can be used to improve the temporal resolution and also acquire multiple slices in a shorter scan.

2.5.2.2 Recovery of prospectively undersampled RT cardiac dataset

The multi-slice free-breathing highly undersampled cardiac dataset described in Section 2.4.4.3 was reconstructed using different methods, as illustrated in Fig 2.6. The techniques used for reconstruction in the different rows along with their reconstruction times are: (a) kt-LR (7.5 hrs) (b) temporal TV (4.7 hrs) (c) PSF (4 mins) (d) ℓ_2-SToRM (24 mins) (e) ℓ_1-SToRM (4.9 hrs). The temporal intensity profile along a vertical cut of the image frames (given by the red dotted line in Fig 2.6.(a)) is also shown for each method. The comparisons are only qualitative since the ground truth dataset was not available. We observe that SToRM reduces streaking artifacts and spatial blurring, compared to other state of the art methods. Specifically, we observe that the myocardial borders are well captured, while details such as the papillary muscles are better defined. We also note that while the image frames of the ℓ_1 and ℓ_2-SToRM reconstructions look similar, the temporal intensity profiles of the ℓ_1 formulation appear sharper.

2.5.2.3 Comparison between free-breathing and breath-held cardiac reconstructions

The quality of the reconstructed free-breathing and breath-held cardiac datasets described in Section 2.4.4.4 are compared in Fig 2.7. The breath-held dataset was reconstructed using CG-SENSE [74], while the free breathing dataset was reconstructed using ℓ_2-SToRM. We show the data corresponding to 2 out of the 5 reconstructed slices. The figure shows results from a particular slice of the breath-held dataset and also its best matching slice from the free-breathing dataset; it was difficult to find

perfect matches between the breath-held and the free-breathing acquisitions. Fig 2.7 shows: (a) 3 cardiac phases from the breath-held cine reconstruction and temporal intensity profile along the yellow dotted line. (b) 3 frames from a single cardiac cycle of the free-breathing dataset and temporal intensity profile along a vertical cut (same cut as the breath-held dataset). Note that the breath-held dataset has a few cardiac phases averaged over many cardiac cycles, while the free-breathing dataset consists of several cardiac cycles. Images from the cardiac cycle of the free-breathing reconstructions which best matched the breath-held images are shown here. We observe that the reconstructed dataset is of comparable quality to the breath-held cine datasets.

2.6 Discussion

We proposed a technique to recover signals from under-sampled measurements, assuming that they lie on a low-dimensional manifold embedded in high-dimensional space. Such a model is satisfied by many real-world signals which can be characterized by low-dimensional parameter vectors. Our reconstruction technique was inspired by the Laplacian Eigenmaps algorithm for dimensionality reduction. The technique relies on exploiting the similarities between signals in local neighbourhoods of the manifold. While our results were demonstrated on dynamic MR image reconstruction problems, the technique is general enough to be used for a variety of applications where the signal model is satisfied, particularly for dynamic imaging using other modalities. In such cases, the acquisition scheme needs to be modified accordingly to enable the estimation of the weight matrix. Once this is done reliably, the reconstruction scheme can be extended fairly easily.

For the problem of dynamic MR image reconstruction, the technique estimates the proximity of the images on the manifold using navigator signals, followed by a manifold aware recovery of the images from highly undersampled measurements. The reconstructed image quality was observed to be superior to that achieved by other state-of-the-art ungated reconstruction methods. Moreover, the experiment on the

speech dataset demonstrated that SToRM can recover images in case of repeating frames, irrespective of whether the repetitions are periodic. In fact, the method does not distinguish between periodic and aperiodic changes. The quality of our reconstructed images is quite dependent on the degrees of freedom of the underlying physiological process. If the degrees of freedom is low, then every frame will have a sufficient number of neighbours very similar to it with high probability (provided that our acquisition time is long enough). If the degrees of freedom is high, then many frames may not have any other frames very similar to it, and the recovered frames will be of poor quality. However, in such situations, other model-based reconstruction schemes should also perform poorly due to lack of redundancy in the data.

While the original stable embedding theory deals with random ortho-projectors [22], our empirical comparisons in section 2.5 show that the radial k-space sampling scheme can estimate the neighbourhood of each image frame quite accurately. Moreover, our experiments also show that approximate estimates of the weight matrix (using one radial line of k-space) are often sufficient to ensure good recovery of images. Our experiments also reveal that spiral navigators are more efficient than radial navigators. We used the radial acquisition scheme for ease of implementation on the scanner. We will investigate the utility of spiral navigators in the future, which may translate to improved temporal resolution or reconstruction quality.

The proposed scheme has a few free parameters: (1) σ (2) the number of neighbours (3) λ. The optimal σ value is dependent on the k-space trajectory as well as the number of points. However, we observed that the reconstruction quality is not very sensitive to the exact value of σ. Specifically, changing σ by a factor of 10 does not significantly affect the reconstruction quality. The number of neighbours is a data dependent parameter determined by the degree of redundancy in the dataset. If a sufficient number of similar frames is available for each frame, then a small increase in the number of neighbours will not affect the image quality. However, if the number

of neighbours is made very high, then all the neighbours of a particular frame will not be very similar to it, and the resulting reconstructed image will have motion blur. If the number of neighbours is made very low, then we will have aliasing artefacts. Similarly, the regularization parameter λ is also data dependent.

We show that ℓ_2-SToRM has similarities to the k-t PCA and PSF methods, with the exception that the temporal basis functions are the singular vectors of the Laplacian matrix rather than that of the covariance matrix. These basis functions promote smoother solutions on the manifold, enabling the exploitation of the non-linear dependencies between images. ℓ_1-SToRM is similar to the temporal TV scheme, with the exception that the standard finite difference matrix is replaced by an adaptive finite difference operator; this enables the exploitation of non-local dependencies between images in the dataset. The ℓ_2-SToRM scheme has similarities to the recent work [7]. Specifically, their solution is a clever approximation of our analytic solution in the ℓ_2 setting for the single channel case. Our approach also has conceptual similarities to [88], where the cardiac and respiratory phase information is recovered from the singular vectors of the graph Laplacian. This approach has been inspired by dimensionality reduction methods such as ISOMAP and LLE [75, 86] that are used to embed the data on a manifold to a lower dimensional subspace. [88] identifies the cardiac and respiratory phases from the dimensionality reduced data, followed by explicit motion-resolved binned reconstructions similar to [25]. In contrast, SToRM performs an implicit motion-resolved recovery of the entire RT dataset. In addition, SToRM does not need the explicit identification of individual phases, which is difficult in applications with both cardiac and respiratory motion and require additional pre-processing steps [26, 88]. The estimation of the cardiac and respiratory phases using band-pass filtering as in [25, 26] may be challenging in cases with irregular respiratory motion and arrhythmia. In addition, many applications like speech imaging have no concept of phase equivalent to cardiac and respiratory phases in cardiac

imaging. SToRM extends readily to such applications. The proposed scheme also has conceptual similarities to recent kernel PCA based approaches, introduced to exploit non-linear similarities between image patches. Specifically, [56] learns the basis functions using linear PCA on non-linearly transformed patches from low-resolution images. They then iterate between projecting each non-linearly transformed patch from the high-resolution images to this subspace, and solving for pre-images that satisfy data-consistency. This approach may be seen as a synthesis formulation of ℓ_2-SToRM, when re-engineered for image patches.

2.7 Conclusion

We introduced a novel acquisition and reconstruction scheme for reconstruction of signals from under-sampled measurements. We demonstrated the technique termed SToRM on real-time dynamic MR imaging. The central assumption is that the images in the dynamic dataset are points on a smooth, low dimensional manifold embedded in high dimensional space. We formulated the recovery of the dataset from highly under-sampled measurements as a manifold smoothness regularized optimization problem. The neighbours of each image on the manifold were estimated from the navigator acquisition. SToRM was demonstrated to be useful in accelerating free breathing cardiac imaging and speech imaging, without compromising on image quality and slice coverage. This approach improves the spatio-temporal resolution, while ensuring patient comfort and reducing the total scan time. It can be easily extended to other dynamic imaging applications like liver, bowel and lung imaging.

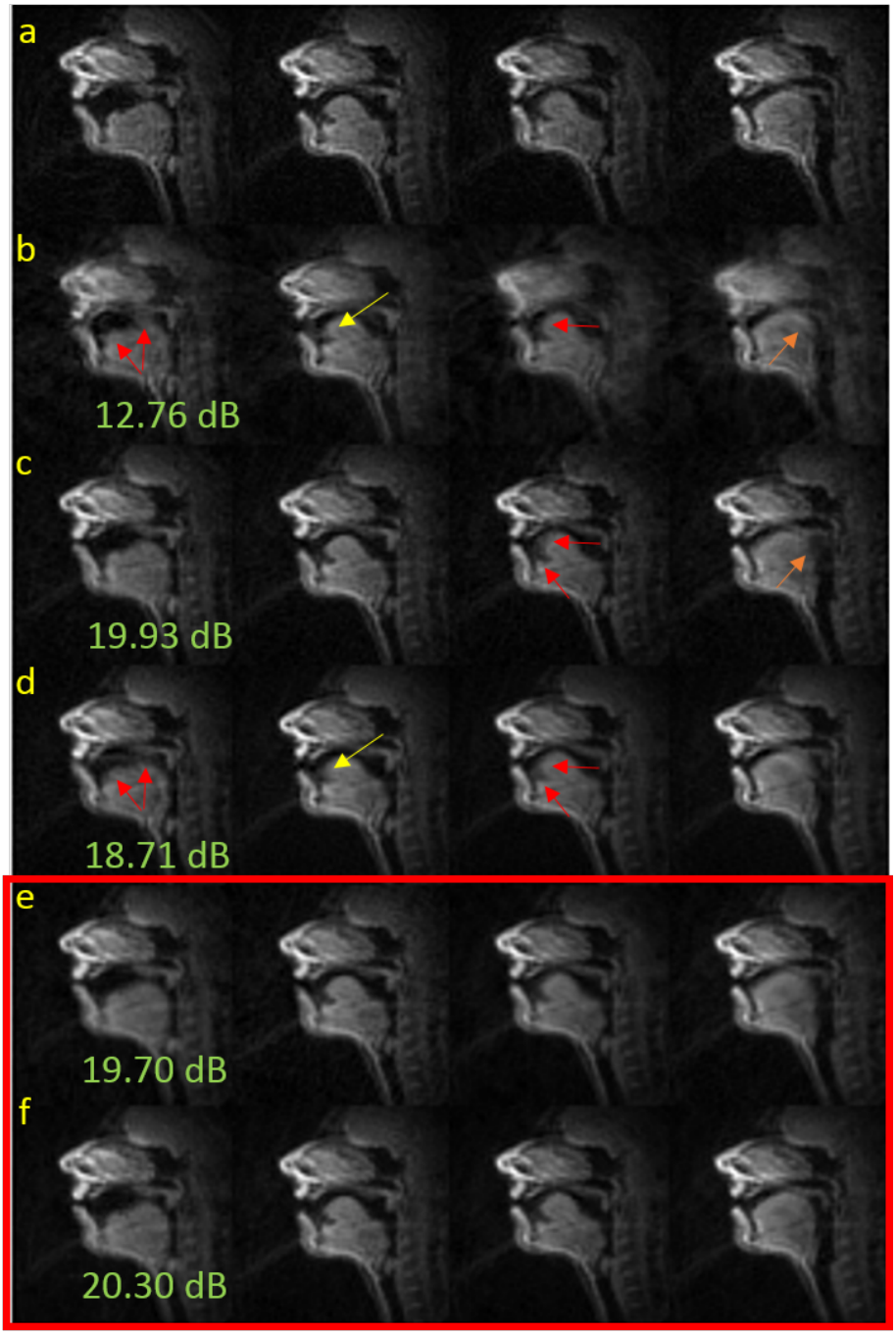


Figure 2.5: Reconstruction of the speech dataset. **(a)** Ground-truth images. The subsequent rows correspond to reconstructions from under-sampled k-space data using **(b)** kt-LR, **(c)** temporal TV, **(d)** PSF, **(e)** ℓ_2-SToRM, and **(f)** ℓ_1-SToRM. The data used for (b) and (c) had a golden angle radial trajectory without navigators. The data used for (d), (e) and (f) had a spiral navigator. The arrows point out artefacts in the images reconstructed by the competing methods, which are not present in the images reconstructed by SToRM.

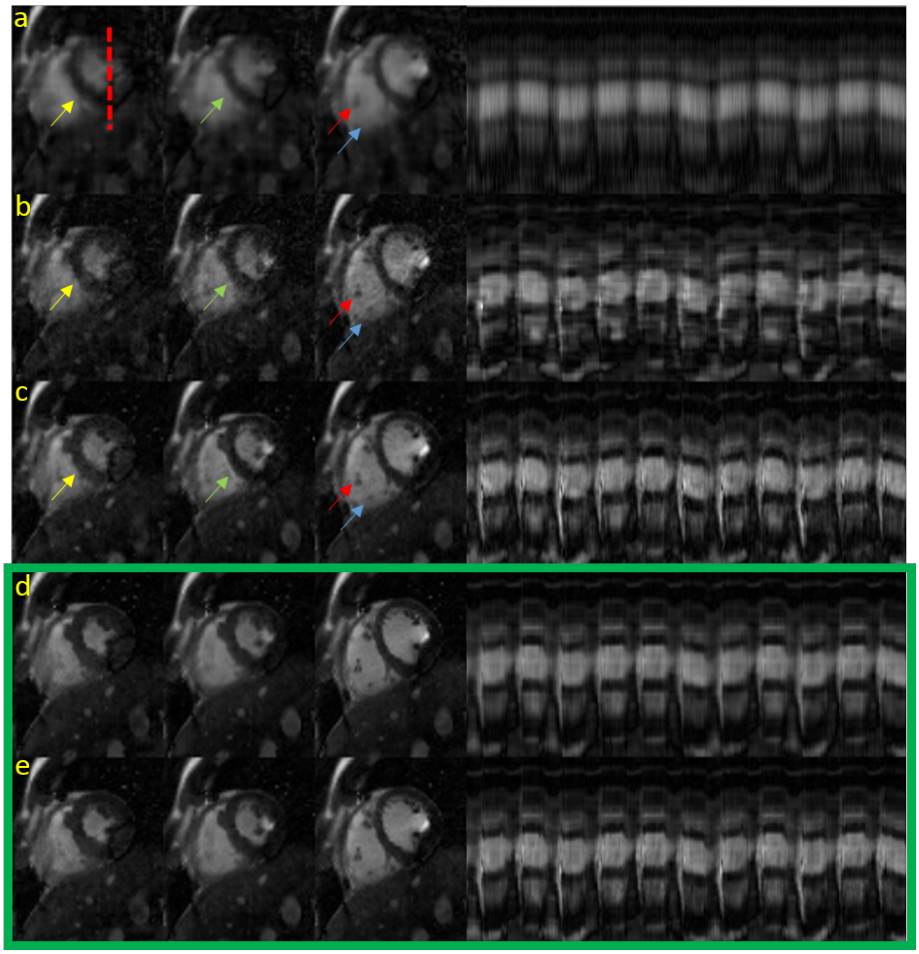


Figure 2.6: Reconstruction of the free-breathing cardiac dataset. Selected image frames and temporal intensity profiles along a vertical cut given by the red dotted line in (a) are shown. The images were reconstructed from under-sampled k-space data using **(a)** kt-LR, **(b)** temporal TV, **(c)** PSF, **(d)** ℓ_2-SToRM, and **(e)** ℓ_1-SToRM. The arrows point out artefacts in the images reconstructed by the competing methods, which are not present in the images reconstructed by SToRM.

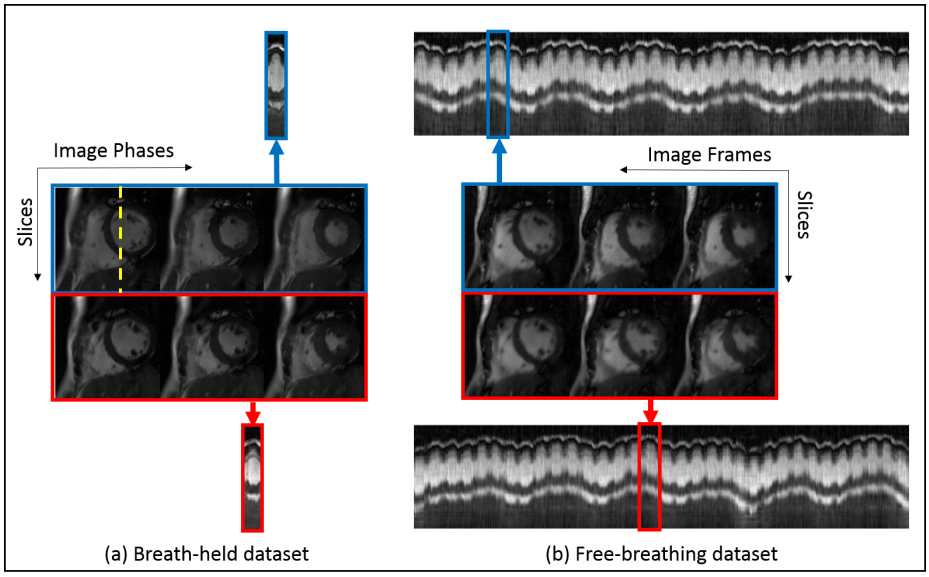


Figure 2.7: Comparison between proposed free-breathing (FB) reconstruction and breath-held (BH) reconstruction. The BH dataset was reconstructed using CG-SENSE. The FB dataset was recovered using ℓ_2-SToRM. Two matching slices from both datasets are shown. The rows represent different slices. **(a)** Images in different cardiac phases from the BH dataset. The voxel profiles along the yellow dotted line are also shown. **(b)** Image frames from a particular cardiac cycle of the FB dataset. The voxel profiles for a few cardiac cycles of the FB dataset are also shown (along the same cut as the BH dataset).

CHAPTER 3
SIGNAL RECOVERY ON SMOOTH CURVES/SURFACES: THEORETICAL GUARANTEES

3.1 Introduction

The main focus of this chapter is to introduce a continuous domain perspective on the recovery of points drawn from a smooth surface in very high dimensions. This work reveals fundamental links between recent advances in superresolution theory [10, 62, 78] and kernel based machine learning methods [79] as well as graph signal processing [82]. We assume that the high dimensional points live on an smooth surface, which is the zero level set of a trigonometric polynomial. This is termed the annihilation relation and it is shown that this relation can be expressed as a weighted linear combination of the exponential features of the point; the dimension of the feature maps is equal to the bandwidth of the polynomial. These properties enable us to determine the sampling conditions, which will guarantee the recovery of the surface from finite number of points. Our analysis also shows that when the bandwidth is overestimated, there are multiple such annihilation relations, suggesting that the exponential feature maps of the points on the surface live in a finite dimensional space. Note that similar non-linear maps are widely used in kernel methods; our results show that these maps can be approximated by a few basis functions, when the points are restricted to a bandlimited surface.

The rank deficient matrix of feature maps translate to a low-rank kernel matrix, computed from the points using a shift invariant kernel such as the Dirichlet function. We minimize the nuclear norm of the feature maps of the points to recover them from noisy data. Since the direct estimation of the surface in higher dimensions suffers from the curse of dimensionality, we use the "kernel trick" to keep the computational complexity manageable. We rely on an iterative reweighted algorithm to recover the denoised points. The resulting algorithm has similarities to iterative non-local methods [29, 41, 54, 55, 93] that are widely used in image processing and graph signal

processing. Specifically, it alternates between the estimation of a graph Laplacian, which specifies the connectivity of the points, and the smoothing of points guided by the graph Laplacian.

This work is built upon our prior work [2,3,60–62,64] and the recent work by Ongie et al., which considered polynomial kernels [63]. Our main focus is to generalize [63] to shift invariant kernels, which are more widely used. We also introduce sampling conditions and algorithms to determine the surface, when the dimension is low. In addition, the iterative algorithm using the kernel trick shows the connections with graph Laplacian based methods used in graph signal processing.

3.2 Exploiting annihilation relations for signal recovery

We assume the point cloud to be supported on a surface in $[-1/2, 1/2]^n$, which is the zero level-set of a bandlimited potential function:

$$\{\mathbf{r} \in \mathbb{R}^n | \psi(\mathbf{r}) = 0\}, \text{ where } \psi(\mathbf{r}) = \sum_{\mathbf{k}\in\Lambda} \mathbf{c}_k \exp(j\ 2\pi \mathbf{k}^T \mathbf{r}) \tag{3.1}$$

Here, $\{\mathbf{c_k} : \mathbf{k} \in \Lambda\}$ is the smallest set of coefficients (minimal set) that satisfies the above relation. $\Lambda \subset \mathbb{Z}^n$ is a set of contiguous locations that indicates the support of the Fourier series coefficients of ψ. Consider an arbitrary point $\mathbf{x}$ on the above surface (3.1). By definition (3.1), we have the annihilation relation $\psi(\mathbf{x}) = \sum_{\mathbf{k}\in\Lambda} \mathbf{c}_k \exp(j\ 2\pi \mathbf{k}^T \mathbf{x}) = 0$. We re-express the annihilation relation as $\mathbf{c}^T \phi_\Lambda(\mathbf{x}) = 0$ using a non-linear mapping $\phi_\Lambda : \mathbb{R}^n \to \mathbb{C}^{|\Lambda|}$:

$$\phi_\Lambda(\mathbf{x}) = \begin{bmatrix} \exp(j\ 2\pi \mathbf{k}_1^T \mathbf{x}) & \dots & \exp(j\ 2\pi \mathbf{k}_{|\Lambda|}^T \mathbf{x}) \end{bmatrix}^T \tag{3.2}$$

This annihilation relation is illustrated in Fig 3.1.

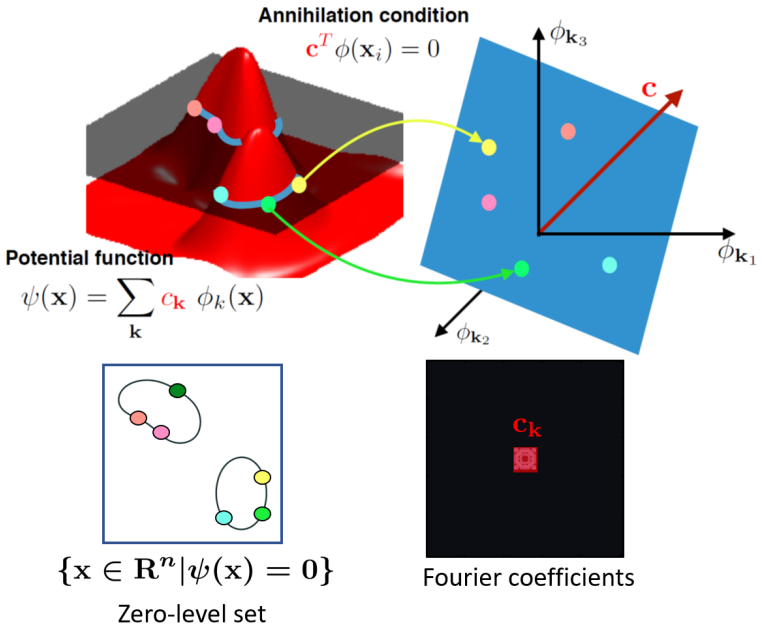


Figure 3.1: Illustration of the annihilation relations in 2-D. We assume that the curve is the zero-level set of a bandlimited function $\psi(\mathbf{x})$. Each point on the curve satisfies $\psi(\mathbf{x}_i) = 0 = \mathbf{c}^T \phi_\Lambda(\mathbf{x}_i)$, which can be seen as an annihilation relation in the non-linear feature space $\phi_\Lambda(\mathbf{x})$. Specifically, the maps of the points lie on a plane orthogonal to $\mathbf{c}$.

3.2.1 Curve recovery: sampling conditions

The annihilation relation introduced in the previous sub-section can be used to estimate the surface, or equivalently $\psi(\mathbf{r})$ from a few number of points. The least square estimation of the coefficients from the data points $\{\mathbf{x}_i : i = 1, \ldots, N\}$ can be posed as the minimization of the criterion:

$$\mathcal{C}(\mathbf{c}) = \sum_{i=1}^{N} \|\psi(\mathbf{x}_i)\|^2 = \mathbf{c}^T \mathbf{Q}_\Lambda \mathbf{c} \tag{3.3}$$

where $\mathbf{Q}_\Lambda = \sum_{i=1}^{N} \phi_\Lambda(\mathbf{x}_i)\phi_\Lambda(\mathbf{x}_i)^T$. The coefficients can be estimated as:

$$\mathbf{c}^* = \arg\min_{\mathbf{c}} \mathbf{c}^T\ \mathbf{Q}_\Lambda\ \mathbf{c} \quad \text{such that} \quad \|\mathbf{c}\|^2 = 1 \tag{3.4}$$

The solution is the minimum eigen vector of $\mathbf{Q}_\Lambda$.

In the remainder of the section, we will restrict our attention to 2-D. We will leave the generalization to higher dimensions for future work. We will now determine the sampling conditions for the perfect recovery of the curve $\psi(\mathbf{x}) = 0$ using (3.4). Specifically, we will determine the minimum number of samples for the successful recovery of the curve, when Λ is a rectangular neighborhood in $\mathbb{Z}^2$ of size $K_1 \times K_2$. In addition, we assume that ψ is the function with the smallest Fourier support (minimal polynomial), whose zeros define the curve. We first focus on the case where Λ is known.

Proposition 3.2.1. *Let $\mathbf{x}_i; i = 1, .., N$ be points on the zero-level set of a band-limited function $\psi(\mathbf{r}), \mathbf{r} \in \mathcal{R}^2$, where the bandwidth of the surface ψ is specified by $|\Lambda| = K_1 \times K_2$ and $\psi(\mathbf{r})$ has J irreducible factors. If N_j points are sampled on the j^{th} irreducible factor, then the curve $\psi(\mathbf{r}) = 0$ can be uniquely recovered by* (3.4)*, when:*

$$N_j > (K_1 + K_2)(K_1^j + K_2^j) \tag{3.5}$$

for $j = 1, \ldots, J$.

Thus, the total number of points required are $N > (K_1+K_2)(K_1+K_2+2(J-1))$. The above proposition is proved in Appendix A.1. We compare this setting with the sampling conditions for the recovery of a piecewise constant image, whose gradients vanish on a bandlimited curve [62]. The minimum number of Fourier measurements required to recover the function there is $|3\Lambda|$; when $K_1 = K_2 = K$, then $3K^2$ complex Fourier samples are required. In contrast, we need $4K^2$ real samples. When the true support Λ is not known, it is a common practice to overestimate it as $\Gamma \supset \Lambda$. In this case, $\mathbf{Q}_\Gamma$ will have multiple null space vectors, as shown below.

Proposition 3.2.2. *We consider the polynomial $\psi(\mathbf{r})$ described in Proposition 1. Let*

$\Lambda \subset \Gamma$ *with* $|\Gamma| = L_1 \times L_2$ *and for* $j = 1, \ldots, J$*:*

$$N_j > (L_1 + L_2)(K_1^j + K_2^j) \tag{3.6}$$

points be sampled on the j^{th} *irreducible factor of* $\psi(\mathbf{r})$*. Then all nullspace vectors* $\mathbf{c}' \overset{\mathcal{F}}{\leftrightarrow} \psi'$ *of the matrix* $\mathbf{Q}_\Gamma$ *will be of the form:*

$$\psi'(\mathbf{r}) = \psi(\mathbf{r})\ \eta(\mathbf{r}) \tag{3.7}$$

where $\eta(\mathbf{x})$ *is an arbitrary function such that* $supp(\mathbf{c}') = \Gamma$*.*

Thus, the total number of points required are $N > (L_1 + L_2)(K_1 + K_2 + 2(J-1))$. The above proposition is proved in Appendix A.2. Since $\psi(\mathbf{x})$ is the common factor of all the annihilating functions, all of them will satisfy $\psi'(\mathbf{x}) = 0$, for any point on the original curve. Depending on the specific η, they will have additional zeros. Hence, the above result provides us a means to compute the original curve, even when the original bandwidth/support of the function is unknown.

We now consider a collection of N points on the curve, stacked into a matrix $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots \mathbf{x}_N]$. Let the feature matrix of size $|\Gamma| \times N$ be denoted by:

$$\Phi_\Gamma(\mathbf{X}) = \Big[\phi_\Gamma(\mathbf{x}_1), \ldots \phi_\Gamma(\mathbf{x}_N)\Big] \tag{3.8}$$

We state a result about the rank of the above feature matrix.

Proposition 3.2.3. *We consider the polynomial* $\psi(\mathbf{r})$ *described in Proposition 1 and* $\Lambda \subset \Gamma$*. Then:*

$$\operatorname{rank}\left(\Phi_\Gamma(\mathbf{X})\right) \leq \underbrace{|\Gamma| - |\Gamma : \Lambda|}_{r} \tag{3.9}$$

with equality if the sampling conditions of Proposition 2 are satisfied.

The above proposition is proved in Appendix A.3. Here, $|\Gamma : \Lambda|$ denotes the

number of valid shifts of the set Λ within Γ as shown in Fig 3.2. Note that as $|\Lambda|$ gets smaller, the number of shifts of it within Γ increases, and hence the rank decreases. The rank of the matrix can be used as a surrogate for the bandwidth of ψ, or equivalently the complexity of the curve. Note that Λ may be an irregular shape in $\mathbb{Z}^n$. For example, if the points lie on a line in $\mathbb{R}^n$, then Λ could be concentrated along a line in $\mathbb{Z}^n$, resulting in a small $|\Lambda|$, even when the number of features in Γ may be considerably high. The low-rank structure of the feature maps can be used to denoise the original points, while the sum of squares function obtained from the nullspace filters can be used to estimate the surface in low-dimensions when (3.6) is satisfied, as illustrated in Fig 3.2.

3.2.2 Recovery of noisy point clouds in high dimensions

The explicit approach of estimating the surface is feasible, when the dimension of the points n is small. However, this approach suffers from the curse of dimensionality. Since the shape of the data, or equivalently the shape of the support Λ is not known, one needs to use a large Γ to ensure that $\Lambda \subset \Gamma$. Note that the dimension of the feature space specified by $|\Gamma|$ grows exponentially with n, making this approach impractical in applications involving point clouds of images or patches.

We hence rely on the right nullspace relations to recover the points from their noisy and undersampled measurements. Specifically, we are interested in the null space relations

$$\underbrace{\Phi_\Gamma(\mathbf{X})^H \Phi_\Gamma(\mathbf{X})}_{\mathbf{K}^\Gamma} \mathbf{v}_i = \mathbf{0} \tag{3.10}$$

where the entries of the $|N| \times |N|$ Gram matrix $\mathbf{K}_\Gamma$ are

$$(\mathbf{K}_\Gamma)_{i,j} = \phi_\Gamma(\mathbf{x}_i)^H \phi_\Gamma(\mathbf{x}_j) = \underbrace{\sum_{\mathbf{k}\in\Gamma} \exp\left(j\ 2\pi \mathbf{k}^T (\mathbf{x}_j - \mathbf{x}_i)\right)}_{\kappa_\Gamma(\mathbf{x}_j - \mathbf{x}_i)} \tag{3.11}$$

The function $\kappa_\Gamma(\mathbf{r})$ in (3.11) is shift invariant and is dependent on the shape of Γ. For example, when Γ is a centered cube in $\mathbf{R}^n$, $\kappa_\Gamma(\mathbf{r})$ is a Dirichlet function. The kernel matrix satisfies $\text{rank}(\mathbf{K}_\Gamma) \leq r$, where r is given by (3.9).

3.2.3 Dirichlet and Gaussian surface representation

The bandlimited function $\psi(\mathbf{r})$ in (3.1) can equivalently be expressed as:

$$\psi(\mathbf{r}) = \sum_{\mathbf{l} \in \Gamma^c} d_\mathbf{l} \; \varphi_\Gamma(\mathbf{r} - \mathbf{l}) \tag{3.12}$$

where $\varphi_\Gamma(\mathbf{x})$ is the Dirichlet function dependent on Γ and Γ^c is the set of sampled locations on the curve. Using reciprocity, the non-linear maps in this case can be shown to be:

$$\phi_\Gamma(\mathbf{x}) = \begin{bmatrix} \varphi_\Gamma(\mathbf{x} - \mathbf{x}_1) & \dots & \varphi_\Gamma(\mathbf{x} - \mathbf{x}_{|\Gamma^c|}) \end{bmatrix}^T \tag{3.13}$$

Since the implicit curve is the zero level set of a linear combination of Dirichlet functions, it may be highly oscillatory. An alternative would be to use a level set expansion in terms of weighted exponentials $\exp(-\pi^2\sigma^2\frac{\|\mathbf{k}\|^2}{2}).\exp(j2\pi\mathbf{k}^T\mathbf{r})$, which could give smoother surfaces. In this case κ_Γ approaches a periodized Gaussian function, as $\Gamma \to \mathbb{Z}^n$, and the Gaussian kernel matrix $\mathbf{K}_\Gamma$ is theoretically full rank. However, we observe that the Fourier series coefficients of a Gaussian function can be safely approximated to be zero outside $|\mathbf{k}| < 3/\pi\sigma$, which translates to $|\Lambda| \approx \left(\frac{6}{\pi\sigma}\right)^n$; i.e., the rank will be small for high values of σ.

3.2.4 Denoising using nuclear norm minimization

We rely on the low rank structure of the kernel matrix $\mathbf{K}$ to recover the noisy points. Specifically, with the addition of noise, the points deviate from the zero set of ψ. A high bandwidth potential function is needed to represent the noisy surface. We propose to use the nuclear norm of the feature matrix as a regularizer in the recovery

of the points from noisy measurements:

$$\mathbf{X}^* = \arg\min_{\mathbf{X}} \|\mathbf{X} - \mathbf{Y}\|^2 + \lambda \|\mathbf{\Phi}(\mathbf{X})\|_* \tag{3.14}$$

We use the IRLS algorithm, where $\mathbf{X}$ is updated as:

$$\mathbf{X}^{(n)} = \arg\min_{\mathbf{X}} \|\mathbf{X} - \mathbf{Y}\|^2 + \lambda \text{ trace}[\mathbf{K}(\mathbf{X})\mathbf{Q}^{(n)}] \tag{3.15}$$

and $\mathbf{Q}^{(n)} = [\mathbf{K}(\mathbf{X}^{(n-1)}) + \gamma^{(n)}\mathbf{I}]^{-\frac{1}{2}}$. Note that the solution for (3.15) involves a system of non-linear equations. Instead, we use gradient linearization to simplify our computations, where $\mathbf{K}(\mathbf{X})$ is a Gaussian kernel matrix:

$$\mathbf{X}^{(n)} = \arg\min_{\mathbf{X}} \|\mathbf{X} - \mathbf{Y}\|^2 + \lambda \text{ trace}(\mathbf{X}^T\mathbf{L}^{(n)}\mathbf{X}) \tag{3.16}$$

with $\mathbf{L}^{(n)} = \mathbf{D}^{(n)} - \mathbf{W}^{(n)}$, $\mathbf{D}^{(n)}_{ii} = \sum_j \mathbf{W}^{(n)}_{ij}$, and

$$\mathbf{W}^{(n)} = -\frac{1}{\sigma^2}\mathbf{K}(\mathbf{X}^{(n-1)}) \odot \mathbf{Q}^{(n)} \tag{3.17}$$

We note the equivalence of the above optimization strategy with widely used non-local means and graph optimization schemes. These schemes estimate a Laplacian matrix $\mathbf{L}$, followed by the minimization of the cost function (3.16). These approaches can thus be seen as fitting a smooth bandlimited surface to the point cloud of patches or signals that are assumed to be on the graph.

3.3 Results

We demonstrate propositions 3.2.1 and 3.2.2 through simulations in Fig 3.2. The various notations used are summarized in Fig 3.2 (a). A phase transition plot in (b) illustrates the probability of correct recovery of the curve from few sampled points, under the assumption that the bandwidth of the underlying trigonometric polynomial

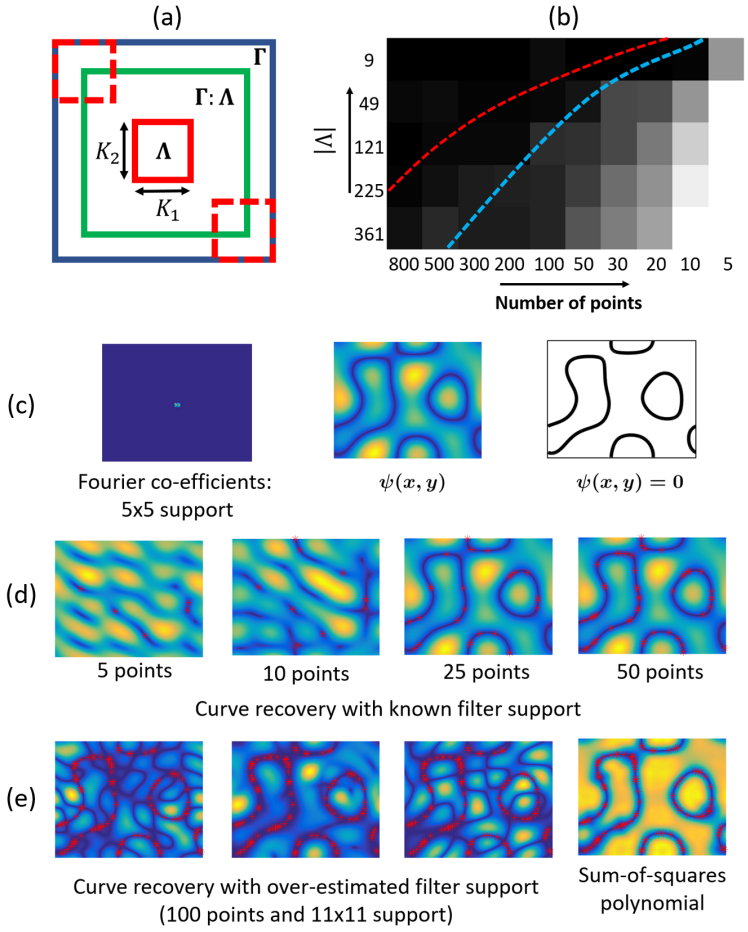


Figure 3.2: Illustration of sampling conditions: The Fourier support Λ of the minimal function ψ, the overestimated support Γ used to evaluate the maps, and the possible shifts of Λ in Γ denoted by $\Gamma : \Lambda$ are shown in (a). In (b), we show a phase transition plot for recovery using known Fourier support, where the red curve is the one predicted by the theory, and the blue curve is for $N = |\Lambda|$. Here, black indicates perfect recovery and white denotes poor recovery. (c) shows an example of a trigonometric polynomial with 5×5 Fourier support, along with its zero-level set. (d) shows the recovery of the curve in (c) from its samples denoted by red points. This experiment assumes that the size of the Fourier support is known. (e) shows the case where the support size was unknown and we assumed Γ to be a 11×11 region. The sum of square of several null space filters uniquely identifies the curve.

is known. A particular example of a curve is shown in (c). Its recovery from sampled points are shown under the assumption of known size of the Fourier support in (d),

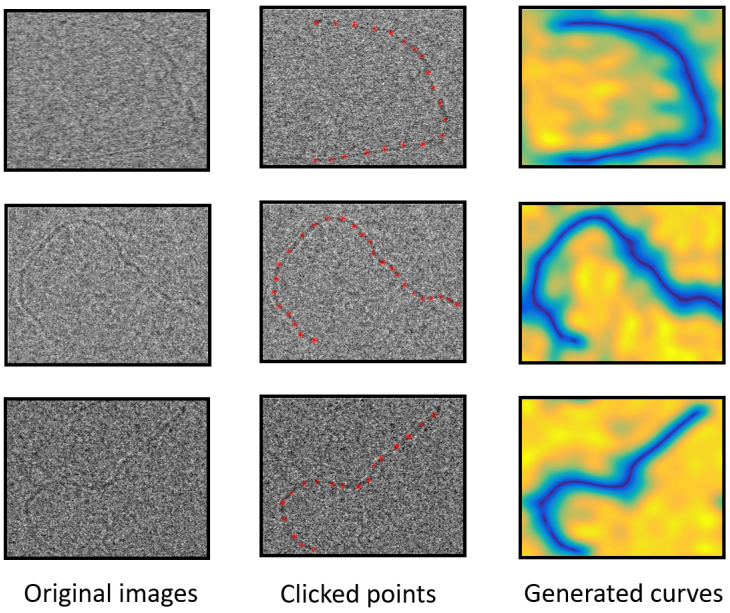


Figure 3.3: Recovery of DNA filaments from few clicked points. The first column shows 3 noisy cryo-electron microscopy images where the DNA filaments are very faintly visible. The second column shows a few points in red that were manually clicked on the noisy images. The third column shows the recovered curves from the clicked points.

and over-estimated size in (e).

We apply the technique of recovering curves from a few sampled points to the problem of DNA filament reconstruction, as shown in Fig 3.3. In this problem, very noisy cryo-electron microscopy images were available where the DNA filaments were visible very faintly. A few points were manually clicked on the filaments. From these few points, the whole filament was recovered. We also demonstrate the utility of (3.14) in a simple 2-D denoising example in Fig 3.4. It is observed that after a few IRLS iterations, the original denoised points are recovered.

3.4 Discussion

We studied the problem of reconstruction of curves/surfaces modelled as the zero-level set of a trigonometric polynomial. In the case of low-dimensional signals, we

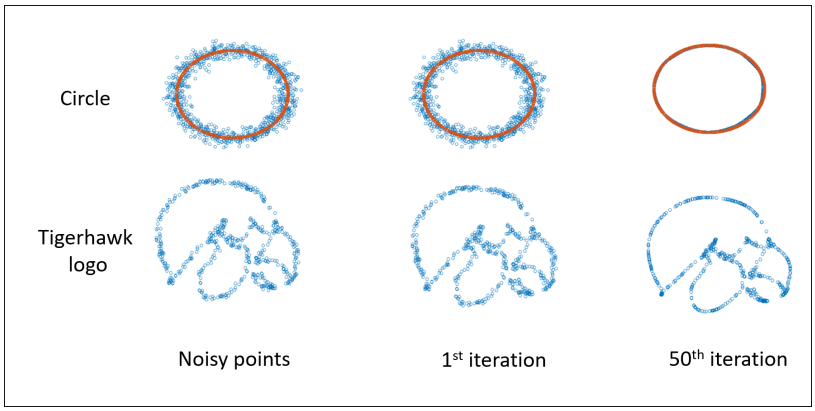


Figure 3.4: Illustration of denoising of 2-D points on a curve using (3.14): The first, second and third columns shows the noisy data, the first iteration of (3.15), and the 50^{th} iterate respectively. Note that the kernel low-rank algorithm provides good recovery of the points with 50 iterations.

consider the case where a few points are sampled on the curve and provide sampling theorems for the perfect recovery of the curves. The number of required points depends on the bandwidth of the underlying trigonometric polynomial. However, we do not consider the effect of noise. Since our technique relies on the detection of the null-space of a large feature matrix, we expect that it may be highly sensitive to noise. This will be studied in future work. Moreover, our theoretical guarantees have also been derived only for 2-D curves, and their extension to higher dimensions also needs to be studied in detail. Our experimental results on the problem of recovering DNA filaments also requires a large number of points to be clicked for perfect recovery. We plan to investigate techniques to reduce this sampling requirement in the future, by studying the effect on the location of the samples on the recovery guarantees.

We study the problem of recovering signals from noisy/under-sampled measurements under the assumption that they satisfy our model. For this purpose, we solve an optimization problem where the regularizer is the nuclear norm of the feature matrix. The feature matrix that is computed depends on the dimension of the signal, and thus for high dimensional signals, this is not memory efficient. For such signals,

we compute the Gram matrix of the feature matrix, whose size is independent of the dimensionality of the signals. We show that the IRLS iterations to solve the proposed scheme can be performed using the Gram matrix alone, without the need for computing the large feature matrix. Our signal model thus provides a basis for several machine learning algorithms which assume that a high-dimensional mapping of the data results is a low-rank matrix.

3.5 Conclusion

We introduced a continuous domain framework for the recovery of points on a bandlimited surface. We show that the exponential maps of the points lie in a lower dimensional subspace, which translates to a kernel matrix that is low-rank. We introduce a nuclear norm minimization algorithm to recover the points. The proposed framework connects the continuous domain surface recovery problem with kernel methods and approaches in graph signal processing. We apply our developed techniques to recover curves from few samples in simulated examples, and also in the context of DNA strand reconstruction. We also demonstrate the scheme on the recovery of points lying on curves from their noisy measurements.

CHAPTER 4
RECOVERY OF CURVES/SURFACES: APPLICATION TO DYNAMIC MRI

4.1 Introduction

We had introduced the SToRM [68] framework for dynamic image reconstruction in chapter 2, which assumes that the images in the free-breathing MRI dataset lie on a smooth and low-dimensional manifold, parameterized by a few variables (e.g. cardiac & respiratory phases). The acquisition scheme relies on navigator radial spokes, which are used to compute the graph Laplacian matrix that captures the structure of the manifold. Conceptually similar manifold models have been proposed by other groups [7, 14, 88]. We, as well as others [7, 88], have relied on the widely used exponential kernel to evaluate the Laplacian entries. To reduce oversmoothing, the entries were then truncated to keep the number of neighbours (degree) of each node fixed, resulting in a regular graph. Note that in practice, we do not have much control on the sampling of the manifold; some manifold neigborhoods are oversampled, while some others are not as well sampled; the use of a regular graph to capture its structure may result in a tradeoff between oversmoothing of poorly sampled regions and good performance in well-sampled regions. We observe that the image quality is quite sensitive to the choice of the node degree. Another challenge with the SToRM algorithm is the need to reconstruct and store the entire dataset (around 1000 frames), which makes the algorithm memory demanding and computationally expensive, and restricts the eventual extension to 3-D applications.

We now propose to use a kernel low-rank formulation for the recovery of dynamic imaging data from undersampled measurements. This approach reconciles SToRM and related approaches [7, 14, 68, 88] with previous kernel low-rank methods [56] that rely on explicit mapping of the data to non-linear features; the explicit approach [56] is restricted to low dimensional signals such as patches or voxel time profiles because of the curse of dimensionality. We model the images as high dimensional points on

a smooth surface/curve, which is represented as the zero level-set of a band-limited function. Under this assumption, feature maps of the images lie on a low-dimensional subspace. We note that previous methods [56] made the assumption of a low-rank kernel matrix, without specifying the underlying model on the images. Since the feature maps are low dimensional, we recover the points from their missing entries using a nuclear norm penalty on their feature maps. The direct implementation of the approach would involve the lifting of the images to high dimensional feature maps, projection to lower-dimensional subspaces, followed by back-projection of the feature maps to images as in [56]; this approach, which is conceptually similar to structured low-rank methods that rely on lifting [30, 38, 60, 64], is prohibitive from a computational and memory perspective when the manifold structure of large images are to be considered. In addition, analytical back-projection steps as in [56] are not available for many feature maps of practical relevance. Motivated by [63], we propose an iteratively reweighted least square (IRLS) algorithm with gradient linearization to directly solve the nuclear norm minimization scheme. This approach does not require the explicit lifting and hence is considerably efficient. IRLS algorithms typically alternate between the estimation of a null-space matrix and a quadratic subproblem, where the penalty term is the energy of the projection to the null-space. In our setting, we alternate between the estimation of a Laplacian-like matrix from the current set of images, and a quadratic SToRM-like subproblem involving the Laplacian-like matrix.

The above link with the proposed kernel low-rank algorithm enables us to further improve the performance of SToRM. To make the recovery from undersampled data well-posed and to further reduce computational complexity, we propose to pre-estimate the Laplacian matrix from k-space navigators; this approach is motivated by similar approaches in low-rank regularization [30, 45, 62, 81, 94]. We estimate the Laplacian matrix from the navigators using an iterative reweighted algorithm. This is a more systematic approach compared to the SToRM approach of using exponential

maps, followed by truncating the neighbours. To further reduce the computational complexity and memory demand of SToRM by an order of magnitude, we approximate the Laplacian matrix by a few of its eigen vectors. The eigen vectors of the Laplacian are termed as Fourier exponentials on the manifold/graph [65]. Instead of reconstructing the entire dataset, we propose to only recover the coefficients of the Laplacian basis functions. Since the framework is an improvement over SToRM using bandlimited modelling of the manifold, we refer to the proposed scheme as b-SToRM. We validate b-SToRM on nine adult congenital heart disease patients with different imaging views, as an add-on to the routine contrast enhanced cardiac MRI study. We study the impact of patient motion, reduced number of navigators, and reduced acquisition time on the algorithm. We also demonstrate that the reconstructed images can be sorted into respiratory and cardiac phases using the eigen-vectors of the estimated Laplacian matrix, facilitating the easy visualization of the data.

4.2 Proposed scheme

We propose to use the low-rank property of the feature matrix to recover the images from the undersampled measurements:

$$\mathbf{X}^* = \arg\min_{\mathbf{X}} \|\mathcal{A}(\mathbf{X}) - \mathbf{B}\|_F^2 + \lambda \ \|\Phi(\mathbf{X})\|_*, \tag{4.1}$$

where $\|\cdot\|_*$ denotes the nuclear norm and $\Phi(\mathbf{X})$ denotes a matrix whose columns are the non-linear maps of the columns of $\mathbf{X}$ (corresponding to different frames). Note that this formulation is similar to structured low-rank methods [30,31,44,60,62,64,81], where the low-rank property of a matrix, whose entries are dependent on the original signal, is exploited. The main difference is that the lifted matrix is now dependent on $\mathbf{X}$ by a non-linear relation, as opposed to linear lifting operators in the classical structured low-rank settings. Note that the above formulation simplifies to low-rank recovery, when $\Phi = \mathcal{I}$, which is the identity map.

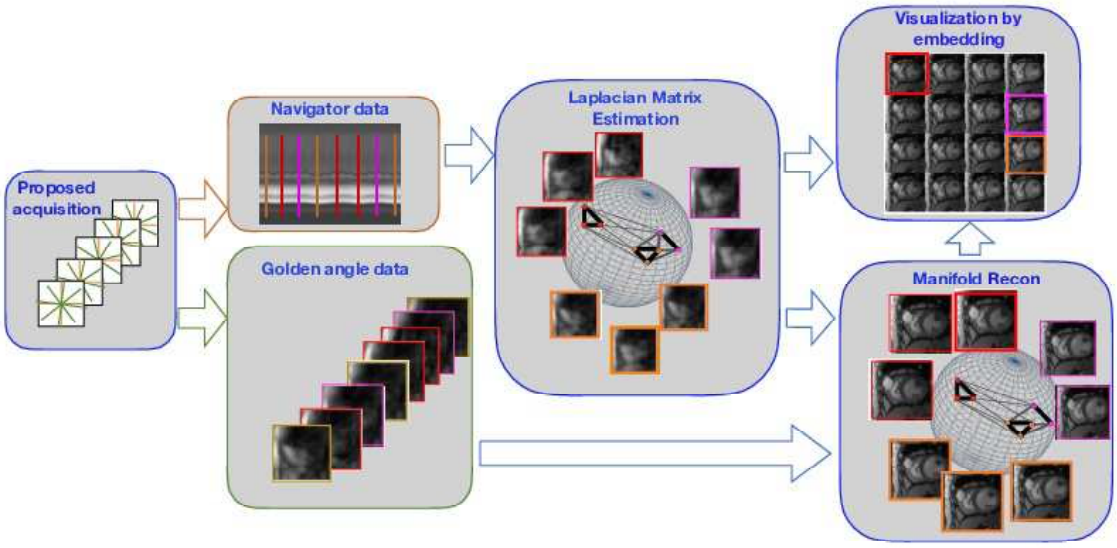


Figure 4.1: Outline of b-SToRM. The free breathing and ungated data is acquired using a navigated golden angle acquisition scheme. We estimate the Laplacian matrix from navigator data using the kernel low-rank model. The entries of the Laplacian matrix specify the connectivity of the points on the manifold, with larger weights between similar frames in the dataset. The manifold is illustrated by the sphere, while the connectivity of the points are denoted by lines whose thickness is indicative of proximity on the manifold. Note that neighbouring frames on the manifold may be well separated in acquisition time. The bandlimited manifold recovery scheme uses the Laplacian matrix to recover the images from the acquired k-space measurements. The Laplacian matrix also facilitates the easy visualization of the data.

This low-rank formulation has conceptual similarities to the approach in [56], where the low-rank structure of pixel intensity profiles that are considerably smaller in dimensions than the images in our setting are considered. In addition, we consider shift invariant kernels unlike the polynomial setting in [56]. Note that the dimension of the feature matrix is even higher than the dimension of the large dynamical imaging dataset $\mathbf{X}$. Hence, the direct approach of lifting the signals, followed by projection to a subspace, and backprojection as in [56] is not feasible in our setting. We hence propose to use the iterative reweighted least squares (IRLS) algorithm [18].

The IRLS algorithm relies on the property:

$$\|\mathbf{Y}\|_* = \operatorname{tr}\left[\mathbf{Y}^*\mathbf{Y}\underbrace{(\mathbf{Y}^*\mathbf{Y})^{-\frac{1}{2}}}_{\mathbf{P}}\right] = \left\|\mathbf{Y}\sqrt{\mathbf{P}}\right\|_F^2 \tag{4.2}$$

to realize an algorithm which alternates between the update of $\mathbf{P} = (\mathbf{Y}^*\mathbf{Y})^{-\frac{1}{2}}$ and the minimization of the quadratic cost function with penalty $\|\mathbf{Y}\sqrt{\mathbf{P}}\|_F^2$. Applying the IRLS algorithm to (4.2), we obtain the following iterations:

$$\mathbf{X}_{n+1} = \arg\min_{\mathbf{X}} \underbrace{\|\mathcal{A}(\mathbf{X}) - \mathbf{B}\|_F^2 + \lambda \operatorname{trace}\left(\mathcal{K}(\mathbf{X})\mathbf{P}_{\mathrm{n}}\right)}_{\mathcal{C}} \tag{4.3}$$

$$\text{where} \qquad \mathbf{P}_n = \left[\mathcal{K}(\mathbf{X}_n) + \epsilon^{(n)}\mathbf{I}\right]^{-\frac{1}{2}} \tag{4.4}$$

Here, the $\mathcal{K}(\mathbf{X}) = \Phi(\mathbf{X})^H\Phi(\mathbf{X})$ is the $k \times k$ Gram matrix of $\Phi(\mathbf{X})$ and $\epsilon^{(n)}$ is a small positive constant added to ensure invertibility. We choose $\epsilon^{(n)} = \frac{\epsilon^{(n-1)}}{\eta}$, where $\eta > 1$ is a constant. Note that this matrix can be computed without explicitly evaluating the feature matrix $\Phi(\mathbf{X})$; the use of this property to speed up algorithms is often termed as the kernel trick [79].

The second term on the RHS of (4.3) involves the non-linear map Φ. Motivated by [63], we focus on the gradient of (4.3) with respect to $\mathbf{X}$. The gradient of the objective function with respect to the i^{th} image $\mathbf{X}_i$ is given by:

$$\nabla_{\mathbf{X}_i}\mathcal{C} = 2\mathbf{A}_i^H\left(\mathbf{A}_i\mathbf{X}_i - \mathbf{b}_i\right) + 2\lambda\sum_j \nabla_{\mathbf{X}_i}[\mathcal{K}(\mathbf{X}_i)]_{ij}\ \mathbf{P}_{ij}^{(n)}$$

When $\mathcal{K}$ is a Gaussian kernel, we can simplify $\nabla_{\mathbf{X}_i}[\mathcal{K}(\mathbf{X}_i)]_{ij}\ \mathbf{P}_{ij}^{(n)} = w_{ij}^{(n)}(\mathbf{X}_i - \mathbf{X}_j)$,

where $w_{ij}^{(n)}$ is the $(i,j)^{th}$ entry of the matrix:

$$\mathbf{W}^{(n)} = -\frac{1}{\sigma^2}\, \mathcal{K}\left(\mathbf{X}^{(n)}\right) \odot \mathbf{P}^{(n)}. \tag{4.5}$$

Here, $\odot$ indicates the point-wise multiplication of two matrices. In matrix form, we thus have:

$$\nabla_{\mathbf{X}} \mathcal{C} = 2\mathcal{A}^H \left(\mathcal{A}(\mathbf{X}) - \mathbf{b}\right) + 2\lambda\, \mathbf{X}\mathbf{L}^{(n)}, \tag{4.6}$$

where

$$\mathbf{L}^{(n)} = \mathbf{D}^{(n)} - \mathbf{W}^{(n)}, \tag{4.7}$$

and $\mathbf{D}^{(n)}$ is a diagonal matrix with elements defined as $\mathbf{D}_{ii}^{(n)} = \sum_j \mathbf{W}_{ij}^{(n)}$. The steepest descent update of (4.3) is given by:

$$\mathbf{X}_{n+1} = \mathbf{X}_n - \gamma_n \left(2\mathcal{A}^H \left(\mathcal{A}(\mathbf{X}) - \mathbf{b}\right) + 2\lambda\, \mathbf{X}\mathbf{L}^{(n)}\right) \tag{4.8}$$

4.2.1 Relation to SToRM regularization

We note that (4.8) can also be viewed as the steepest descent update of the quadratic cost function:

$$\mathbf{X}_{n+1} = \arg\min_{\mathbf{X}} \|\mathcal{A}(\mathbf{X}) - \mathbf{B}\|_F^2 + \lambda\ \mathrm{tr}\left(\mathbf{X}\mathbf{L}^{(\mathrm{n})}\mathbf{X}^{\mathrm{H}}\right), \tag{4.9}$$

which is essentially the main cost function solved in SToRM [68]. Thus, the IRLS scheme can also be interpreted as an algorithm that alternates between SToRM and an update of $\mathbf{L}$ using (4.7) and (4.5). This result shows the link between kernel low-rank regularization and SToRM. The main difference between the methods is that the matrix $\mathbf{W}$ is derived from the current iterate using a fundamentally different formula as in (4.5), as opposed to its estimation from the navigators using (2.16), followed by truncation to obtain a regular graph.

The computational complexity and memory demand of the above iterative reweighted algorithm is expected to be high, especially since the data involving $500-1000$ frames is heavily undersampled in $k-t$ space. Two-step algorithms have been introduced by several researchers in low-rank regularization to reduce the computational complexity and memory demand of structured low-rank algorithms. These methods estimate the signal subspace (or equivalently the null-space) from fully sampled k-space subregions or navigator data, which is then used to solve for the signal. In our prior work in the context of structured low-rank matrix regularization, we estimated the matrix $\mathbf{P}$ in (4.2) that approximates the null-space of the matrix, which was used to solve for the signal. We now propose a similar strategy, where we estimate the $\mathbf{L}$ matrix in (4.9), to obtain a computationally feasible framework.

4.2.2 Two step recovery using $k-t$ space navigators

We acquire multi-coil $k-t$ space navigators $\mathbf{Z}=\Phi\mathbf{X}$ as described in Section 2.3.1. Since this data is corrupted by noise and subtle subject motion, we propose to estimate $\mathbf{L}$ using kernel low-rank regularization. Specifically, we solve:

$$\mathbf{R}^*=\arg\min_{\mathbf{R}}\|\mathbf{R}-\mathbf{Z}\|_F^2+\lambda\quad\|\Phi\left(\mathbf{R}\right)\|_*. \tag{4.10}$$

Solving the above optimization scheme using IRLS as discussed in Section 4.2, we obtain the alternating algorithm

$$\mathbf{R}^{(n)}=\arg\min_{\mathbf{R}}\|\mathbf{R}-\mathbf{Z}\|_F^2+\lambda\ \ \mathrm{tr}\left(\mathbf{R}\ \mathbf{L}^{(n)}\ \mathbf{R}^H\right), \tag{4.11}$$

where

$$\mathbf{L}^{(n)}=\mathbf{D}^{(n)}-\mathbf{W}^{(n)}. \tag{4.12}$$

Here, $\mathbf{W}^{(n)}=-\frac{1}{\sigma^2}\mathcal{K}(\mathbf{R}^{(n)})\odot\mathbf{P}^{(n)}$, where $\mathbf{P}^{(n)}=\left[\mathcal{K}\left(\mathbf{R}^{(n)}\right)+\epsilon^{(n)}\mathbf{I}\right]^{-\frac{1}{2}}$. Note that the size of $\mathbf{Z}$ is considerably smaller than $\mathbf{X}$; the computational complexity of the above

algorithm to solve (4.10) is significantly lower than (4.9). When the above iterations converge, we use the final $\mathbf{L}$ to recover the image frames from their undersampled measurements.

Our empirical results show that the estimation of the $\mathbf{L}$ matrix as the by-product of the above IRLS scheme is considerably more robust than the use of (2.16). In addition to being more robust to noise and subject motion, this approach do not require us to artificially truncate the weight matrix or restrict the number of neighbours to obtain a regular graph. Note that we do not constrain the degree of the nodes, and hence they can be arbitrary. In our experiments, we observe that the off diagonal entries of $\mathbf{L}$ for any specific row are often small with few significant entries.

4.2.3 Approximation of Laplacian matrix for fast computation

We now propose to use the property of the $\mathbf{L}$ matrix to reduce the computational complexity and memory demand of the algorithm. Denoting the eigen decomposition of the symmetric Laplacian matrix as $\mathbf{L} = \mathbf{V}\Sigma\mathbf{V}^H$, we rewrite the SToRM cost function in (2.17) as:

$$\begin{aligned}\mathbf{X}^* &= \arg\min_{\mathbf{X}} \|\mathcal{A}(\mathbf{X}) - \mathbf{B}\|_F^2 + \lambda \ \mathrm{tr}\left[\underbrace{(\mathbf{XV})}_{\mathbf{U}} \ \boldsymbol{\Sigma} \underbrace{(\mathbf{XV})^H}_{\mathbf{U}^H}\right] \\ &= \arg\min_{\mathbf{X}} \|\mathcal{A}(\mathbf{X}) - \mathbf{B}\|_F^2 + \lambda \ \sum_{i=1}^{k} \sigma_i \left\| \underbrace{\mathbf{X}\,\mathbf{v}_i}_{\mathbf{u}_i} \right\|^2 \end{aligned} \tag{4.13}$$

Here, the columns of $\mathbf{V}$ form an orthonormal temporal basis set and $\mathbf{u}_i$ are the spatial coefficients.

We observe that the eigen values often increase rapidly, if $\mathbf{L}$ is the Laplacian matrix. Hence, the weighted norm in the penalty encourages signals $\mathbf{X}$ that are maximally concentrated along the eigen vectors $\mathbf{v}_i$ with small eigen values; these eigen vectors correspond to smooth signals on the manifold. While this reformulation

was introduced in [68] to show similarity with PSF methods, we did not make use of this property to accelerate the algorithm.

We now observe that in the optimization scheme (4.13) the projections of the recovered signal onto the higher singular vectors are expected to be small. We pick the r smallest eigen vectors of $\mathbf{L}$ to approximate the recovered matrix as:

$$\mathbf{X} = \mathbf{U}_r \mathbf{V}_r^H \tag{4.14}$$

where $\mathbf{U}_r$ is a matrix of r basis images (typically around $r \approx 30$) and $\mathbf{V}_r$ is a matrix of r eigen vectors of L with the smallest eigen values. Thus the optimization problem (2.17) now reduces to:

$$\mathbf{U}^* = \arg\min_{\mathbf{U}} \|\mathcal{A}(\mathbf{U}\mathbf{V}^H) - \mathbf{B}\|_F^2 + \lambda \sum_{i=1}^{r} \sigma_i \|\mathbf{u}_i\|^2 \tag{4.15}$$

We observe $r \approx 30$ is sufficient to approximate (4.13) with high accuracy. Since we only have to recover r coefficient images from the measurements, the optimization problem is an order of magnitude more computationally efficient than (2.17). The outline of our scheme is illustrated in Fig 4.1.

4.2.4 Visualization using manifold embedding

Laplacian eigen-maps rely on the eigen vectors of the Laplacian matrix to embed the manifold to a lower dimensional space. When the signal variation in the dataset is primarily due to cardiac and respiratory motion, the second and third lowest eigen vectors are often representative of the cardiac and respiratory phases. This information may be used to bin the recovered data into respiratory and cardiac phases for visualization as in Fig 4.8, even though we do not use explicit binning for image recovery. This post-processing step can be thought of as a manifold embedding scheme using an improved Laplacian eigen-maps algorithm [5], where the main difference

with [5] is the estimation of the Laplacian.

4.3 Results

Cardiac data was collected in the free-breathing mode from nine patients at the University of Iowa Hospitals and Clinics on a 1.5 T Siemens Aera scanner. The institutional review board at the local institution approved all the in-vivo acquisitions and written consent was obtained from all subjects. A FLASH sequence was used to acquire 10 radial lines per frame out of which 4 were uniform radial navigator lines and 6 were Golden angle lines. The sequence parameters were: TR/TE=4.3/1.92 ms, FOV=300mm, Base resolution=256, Bandwidth=574Hz/pix. 10000 spokes of k-space were collected in 43 s. Data corresponding to two views (two-chamber/short-axis and four-chamber) was collected for each patient, resulting in a total of 18 datasets. We used b-SToRM to reconstruct these datasets. The parameters of the image reconstruction algorithm were manually optimized on one dataset, and kept fixed for the rest of the datasets.

We compare the reconstructions from 2 datasets using our technique to a few other competing methods:

1. PSF scheme [45]: For this method, we estimated the temporal profiles using the navigator signals. The recovery of the spatial coefficients was then posed as a least-squares optimization problem, regularized by the Frobenius norm of the spatial coefficients. The number of basis functions was fixed to 30.

2. SToRM [68]: The SToRM scheme was applied using our default parameter settings for both datasets. The exponential weight matrix was thresholded to retain only 2 neighbours per frame.

3. SToRM with few basis functions: For this method, we estimated the weight matrix as in SToRM and formed the Laplacian matrix corresponding to it. A few eigen-vectors of the Laplacian matrix were retained as the temporal basis

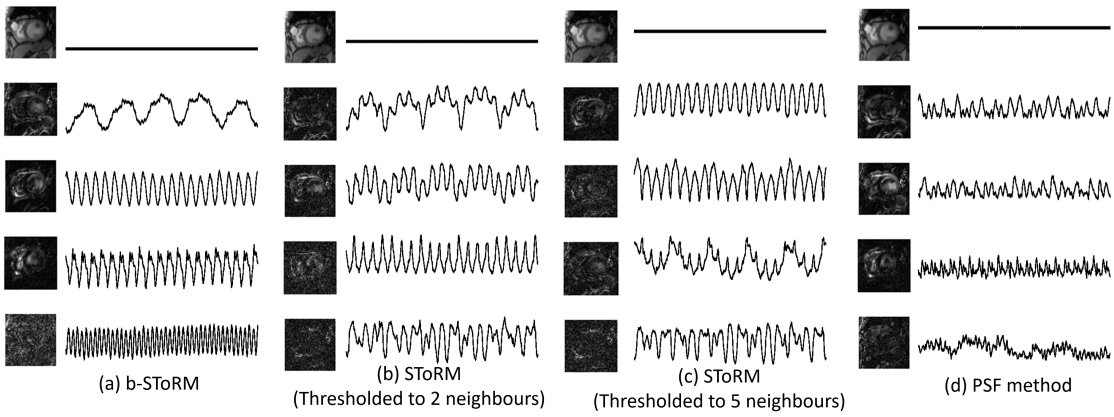


Figure 4.2: Visualization of the basis images and temporal functions. We compare the matrices $\mathbf{U}_r$ and $\mathbf{V}_r$ defined in (4.14) obtained using different methods that employ factorization of the Casorati matrix. (a) corresponds to b-SToRM, while (b) & (c) correspond to the SToRM approach (exponential weight matrix, followed by truncation) of estimating the Laplacian matrix, where 2 and 5 neighbours per node are retained. The temporal basis functions are the eigen vectors $\mathbf{V}$ of the estimated Laplacian matrix with the smallest eigen values. For the PSF scheme, the temporal basis functions are the eigen vectors of the navigator signal matrix with the smallest eigen values. These are shown in (d). It is observed that b-SToRM provides more accurate estimates of cardiac and respiratory motion than the other schemes, thus facilitating the recovery of smooth signals on the manifold. Moreover, by comparing (b) and (c), it is observed that the basis functions are quite sensitive to the choice of the threshold used to compute the SToRM exponential weight matrix.

functions. The spatial co-efficients were then obtained using (4.15). The number of basis functions was fixed to 30.

4. XD-GRASP [25]: We adapted the authors' code that is available online for contrast enhanced liver MRI, to the setting of free-breathing cardiac MRI, using [25] as a guideline. For both datasets, we assumed 10 respiratory phases and 18 cardiac phases, and manually tuned the regularization parameters for best visual quality.

The basis images and temporal profiles obtained using different schemes that utilize the factorization of the Casorati matrix are illustrated in Fig 4.2. We note that the temporal basis functions obtained with the b-SToRM scheme in (a) captures

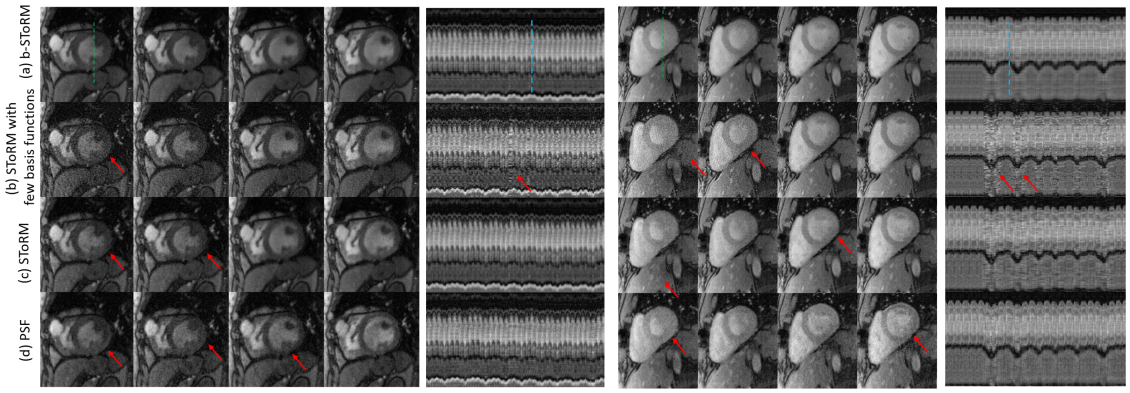


Figure 4.3: Comparison with other methods. Few frames and temporal profiles are shown from two datasets reconstructed using (a) b-STORM (b) STORM using few basis functions (c) STORM [68] (d) PSF scheme [45]. It is observed that b-STORM yields the best overall results, followed by STORM that shows some degradation in image quality indicated by the red arrows. Note that b-STORM also benefits from a speed-up due to the factorization of the Casorati matrix. It is also observed from (b) that using a few basis functions of the STORM Laplacian matrix results in artefacts in the images and the temporal profile. Specifically, the approximation of the STORM Laplacian matrix using few basis functions is poor, which translates to poor recovery. The PSF method also shows some image artefacts as compared to b-STORM, which shows the benefit of the non-linear manifold modeling over subspace approximation. The red arrows in the figure point to artefacts in the images reconstructed using the competing methods.

the physiological components of the motion. Specifically, we observe that the 2^{nd} and 3^{rd} lowest eigen vectors correspond to the respiratory and cardiac motion respectively, while the higher eigen vectors can be thought of as harmonics of the above dynamics. By contrast, the STORM estimates show mixing of the dynamics. The comparison of (b) and (c) show the sensitivity of the estimates to the degree of the regular graph; the approximation of the manifold samples by a regular graph is poor. While the PSF scheme also relies on the factorization of the Casorati matrix, the non-linear manifold model facilitates the better representation of the non-linear dynamics in free-breathing datasets with large respiratory motion.

The b-STORM scheme is compared to other competing methods in Fig 4.3. A few reconstructed images and temporal profiles are shown for (a) b-STORM (b) STORM

with few basis functions (c) SToRM (d) PSF method. It is observed that the images in (a) show less artefacts as compared to the competing methods. In addition, the computational complexity of (a) is significantly lower than (c). The PSF scheme shows some streaking artefacts. In (b), we observe that there are some artefacts in the temporal profile, especially in the challenging dataset to the right which has sudden gasps of breath. This could be because a few eigen-vectors do not capture the physiological motion in this case, or equivalently the approximation of the SToRM Laplacian matrix using few eigen vectors is poor. More frames of this challenging dataset as reconstructed by b-SToRM are shown in Fig 4.4 and Fig 4.5.

We show the comparison of b-SToRM with XD-GRASP in Fig 4.4 (a). Only 4 respiratory and 5 cardiac phases are shown here for better visualization. The reconstructions using b-SToRM are also re-arranged in Fig 4.4 (b) for a direct comparison to (a). For the purpose of re-arranging the frames of the b-SToRM dataset, we used the cardiac and respiratory signals that were estimated using XD-GRASP from the centre k-space temporal profile. It is observed that the images obtained using b-SToRM have less artefacts due to motion and noise, especially in cardiac and respiratory phases which only have a few k-space samples (bottom row). The frames reconstructed using XD-GRASP are also re-arranged to recover a temporal profile. It is observed that the temporal profile is quite noisy and motion is also suppressed, which is due to the discrete segmentation of the frames into phases.

We conduct a few experiments to study the performance of our method in different datasets, and with different acquisition parameters. We study the impact of motion patterns on the reconstructions, using two of the most challenging datasets, with different breathing and cardiac patterns. The datasets in Fig 4.5 have a high amount of respiratory and out-of plane motion, compared to the other datasets that we have collected. The first dataset shows a normal cardiac rate (68 beats/min) accompanied by a very irregular breathing pattern, characterized by several large gasps of breath.

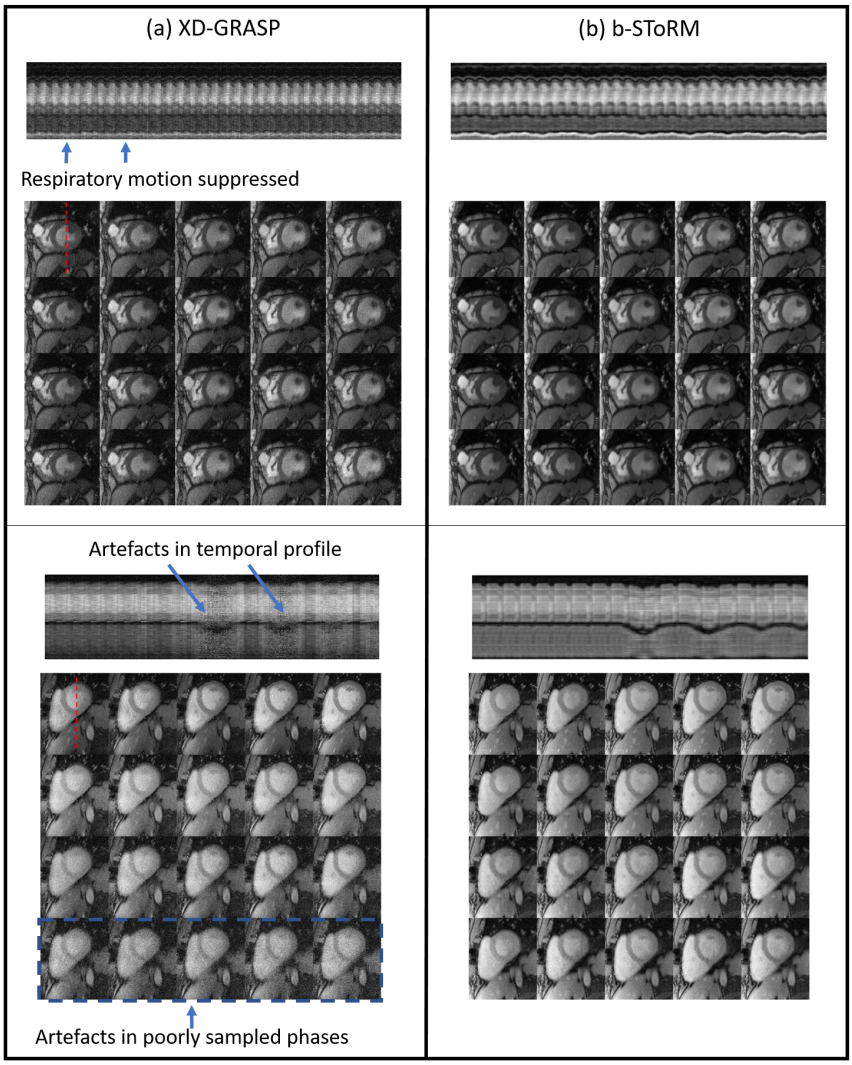


Figure 4.4: Comparison to XD-GRASP: Images corresponding to a few cardiac and respiratory phases reconstructed using XD-GRASP are shown in (a). Since both methods use drastically different reconstruction strategies, we rearrange the images obtained using b-SToRM into respiratory and cardiac phases in (b) for direct comparison to (a). Likewise, the recovered frames of XD-GRASP are also re-arranged to form a temporal profile. It is seen that the images and temporal profiles in (a) have more artefacts as compared to (b). Specifically, it is seen from the temporal profile of (a) that respiratory motion is suppressed. The images in (a) also contain speckle-like artefacts. The image artefacts are more pronounced in the dataset at the bottom where there are sudden gasps of breath, and thus some respiratory phases are very poorly sampled. In comparison, b-SToRM can recover more natural-looking images and temporal profiles.

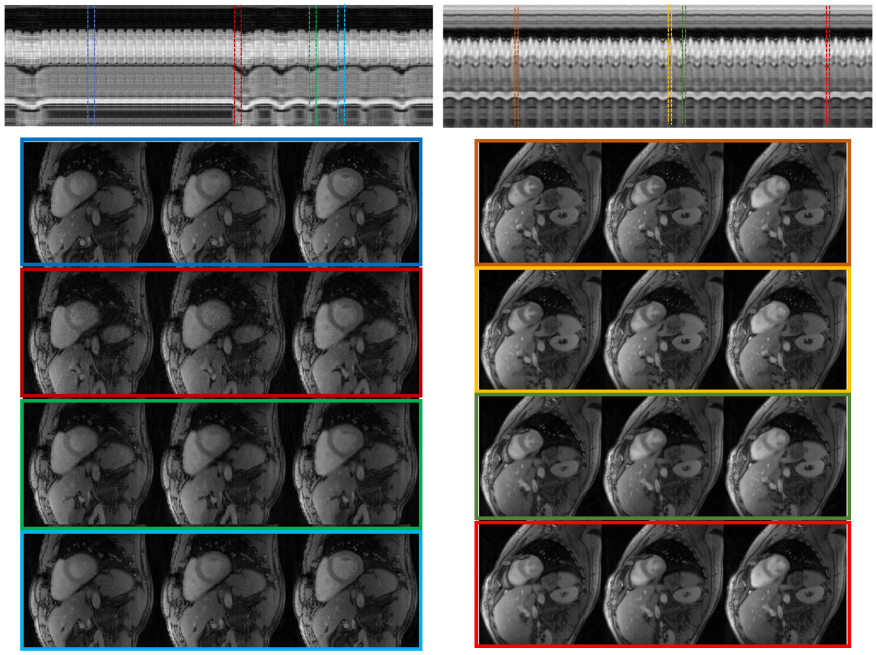


Figure 4.5: Sensitivity of the algorithm to high motion. We illustrate b-SToRM on datasets acquired from two patients with different types of motion. For both datasets, we show a temporal profile for the whole acquisition to give an idea of the amount of breathing and cardiac motion present. We also show a few frames from time points with varying respiratory phase. The dataset on the left has regions with abrupt breathing motion at a few time points. Since these image frames have few similar frames in the dataset (poorly sampled neighbourhood on the manifold), the algorithm results in slightly noisy reconstructions at the time points with high breathing motion (red box). The regions with low respiratory motion (blue and light blue boxes) are recovered well. The dataset on the right shows consistent, but low respiratory motion. By contrast, the heart rate in this patient was high. We observe that b-SToRM is able to produce good quality reconstructions in this case, since all neighbourhoods of the manifold are well sampled.

We show a few reconstructed frames from different time points, at various states of motion. The reconstruction quality is better in presence of less respiratory motion since there are frames similar to it in the dataset; the manifold neighbourhood is well sampled in these neighbourhoods. By contrast, the images are seen to be more noisy in manifold regions that are not well-sampled (red box). The second dataset shows a high cardiac rate (107 beats/min) accompanied by heavy regular breathing (42 breaths/min). We observe that the algorithm is able to reconstruct this case satisfactorily, despite the rapid motion since the manifold is well-sampled.

We also study the effect of the number of navigator lines on the quality of the recovered images using a dataset with a large amount of breathing motion. The main goal is to determine the minimum number of navigator lines per frame to acquire in future studies. For this purpose, we compared the reconstruction using 4 navigator lines to that using only 1 or 2 navigator lines. Two experiments were conducted using 2 navigator lines per frame (corresponding to 0° and 90°) and 1 navigator line per frame (corresponding to 0°) respectively to estimate the weights. For the purpose of reconstruction, we used the full data (6 golden angle lines and 4 navigators). We observe from Fig 4.6 that for both high and low motion regions, there is no degradation in image quality when the number of navigator lines are reduced to two from four. Only using one navigator spoke induces some error, especially for the frames highlighted in green since they have more respiratory motion. This is expected since the approach will only be sensitive to the motion in one direction and not to the direction orthogonal to it. As a result of this experiment, we plan to keep only two navigator lines per frame in the future, and consequently increase the number of golden angle lines to 8 (from 6 in the current acquisition). This should improve image quality by making the sampling patterns between frames more incoherent.

We also study the impact of the acquisition duration on image quality. For 2 datasets with different types of motion patterns, we compare the reconstruction using

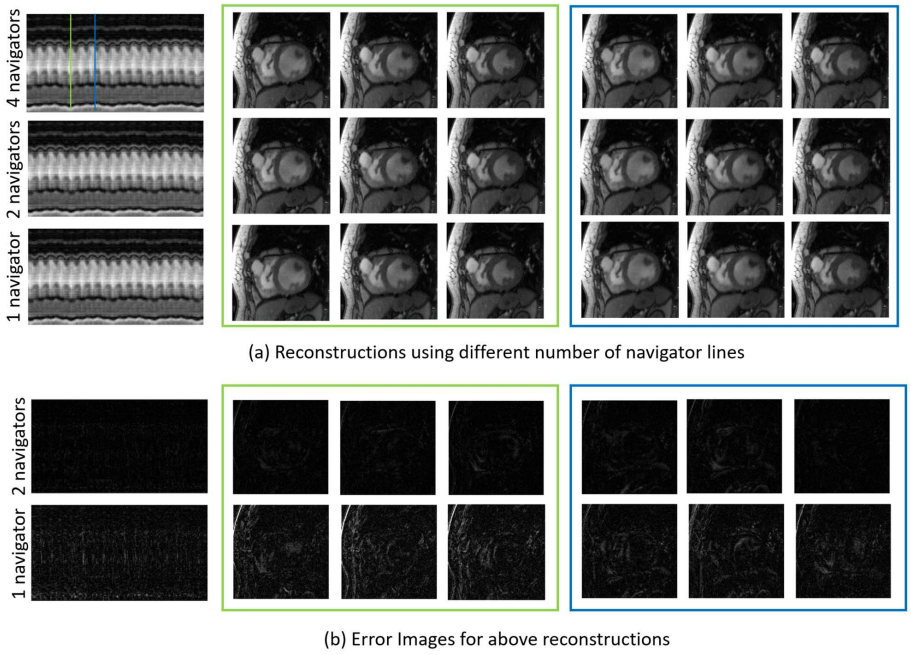


Figure 4.6: Effect of number of navigator lines on the reconstruction quality. We perform an experiment to study the effect of computing the Laplacian matrix $\mathbf{L}$ from different number of navigator lines. For this purpose, we use one of the acquired datasets with 4 navigator lines per frame. We compute the ground-truth $\mathbf{L}$ matrix using all 4 navigators. Next, we also estimate the $\mathbf{L}$ matrix using 2 navigator lines (keeping only the 0° and 90° lines) and 1 navigator line (keeping only the 0° line). We now reconstruct the full data using these three Laplacian matrices, as shown in the figure. We observe that two navigator lines are sufficient to compute the Laplacian matrix reliably. Using one navigator line induces some errors, especially in the frames highlighted in green which are from a time point with higher respiratory motion. As a comparison, note that the error images are in the same scale as those for Fig 4.7.

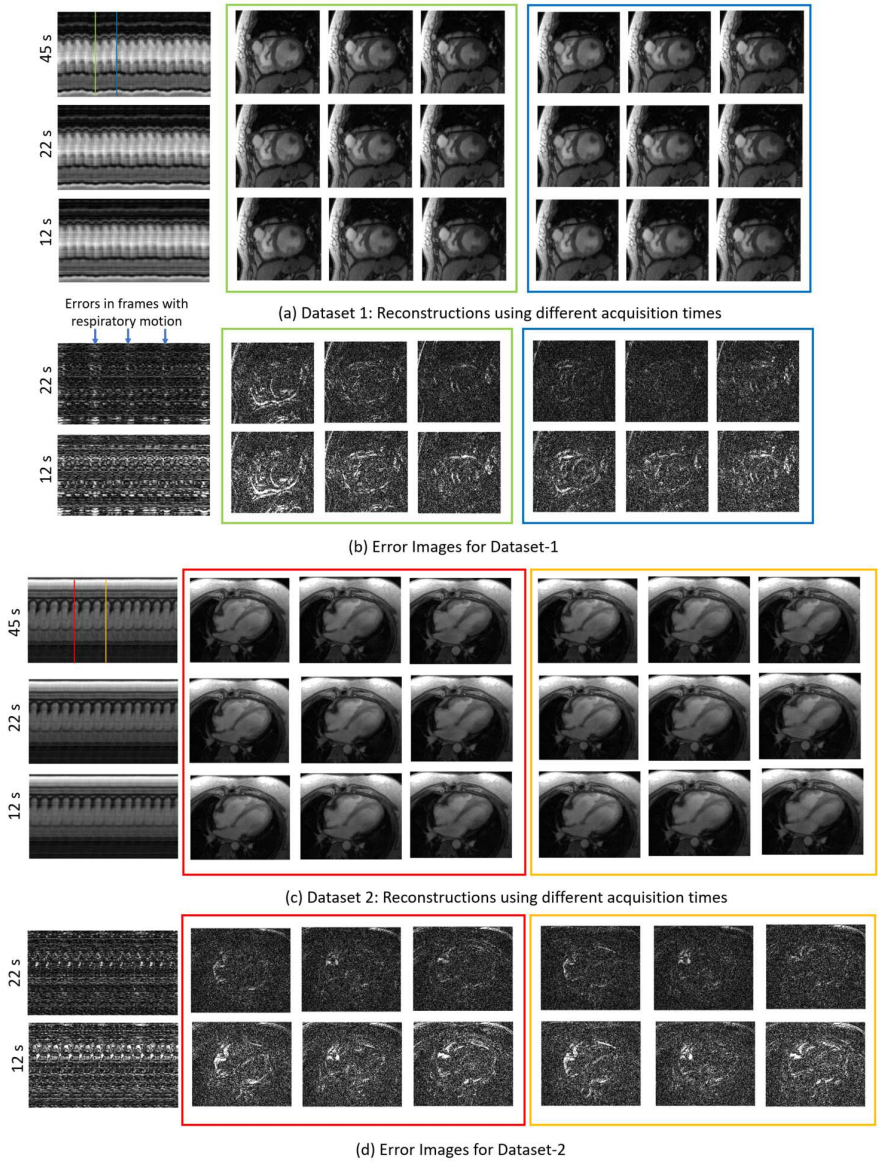


Figure 4.7: Effect of number of frames on the reconstruction quality. We perform an experiment to study the effect of reconstructing the data from a fraction of the time-frames acquired. The original acquisition was 45 seconds long, resulting in 1000 frames. We compare the reconstruction of the 1^{st} 250 frames, using (1) all 1000 frames (2) only 550 frames, i.e. 22 s of acquisition (3) only 350 frames, i.e. 12 s of acquisition. As can be seen from the temporal profiles, Dataset-1 has more respiratory motion than Dataset-2. Consequently, the performance degradation in Dataset-1 is more pronounced with decrease in the number of frames. Moreover, the errors due to decrease in the number of frames is mostly seen in frames with higher respiratory motion, as pointed out by the arrows. As a comparison, note that the error images are in the same scale as those for Fig 4.6.

the entire data, 450 contiguous frames corresponding to 22 s, and also 300 frames corresponding to 12 s. The effect of reducing acquisition time is illustrated in Fig 4.7. The dataset at the top has more breathing motion as compared to the bottom one. We observe that the bottom dataset is robust to decrease in the number of frames; it can be reliably recovered even from 12 seconds of data. The top dataset is more sensitive to reduction in scan time. The green line corresponds to the lowest position of the diaphragm, which is less frequent in the dataset. By contrast, the blue line corresponds to a more frequent frame. The frames around the green line, shown in the green box are more noisy when the scan time is reduced to 12 seconds, compared to the reconstructions within the blue box. We observe negligible errors in both datasets when the acquisition time is reduced to 22s, whereas relatively noisy reconstructions are seen in high motion frames when it is reduced to 12 second acquisition windows. The error images for Fig 4.6 and Fig 4.7 are on the same scale, to illustrate the relative effects of changing the number of navigators and the number of frames.

We demonstrate that the recovered data can be automatically binned into respiratory and cardiac phases using two eigen-vectors of the estimated Laplacian matrix. Thanks to the accurate and robust estimation of the Laplacian matrix, these eigen-vectors accurately represent the respiratory and cardiac motion of the patient over the entire acquisition. Using this information, each image frame can be assigned a bin depending on its respiratory and cardiac phase. Images from each bin can be viewed to find representative members of a particular cardiac or respiratory phase. The results in Fig 4.8 show that the improved Laplacian eigen maps approach facilitates the easy visualization of the data. In general, we observe that the eigen-vectors of the Laplacian matrix with the second and third lowest eigen values correspond to respiratory and cardiac motion. It can be appreciated from Fig 4.2 that such a binning strategy is not possible when the exponential weights are used.

Fig 4.9 demonstrates the potential of b-SToRM to replace clinical breath-held

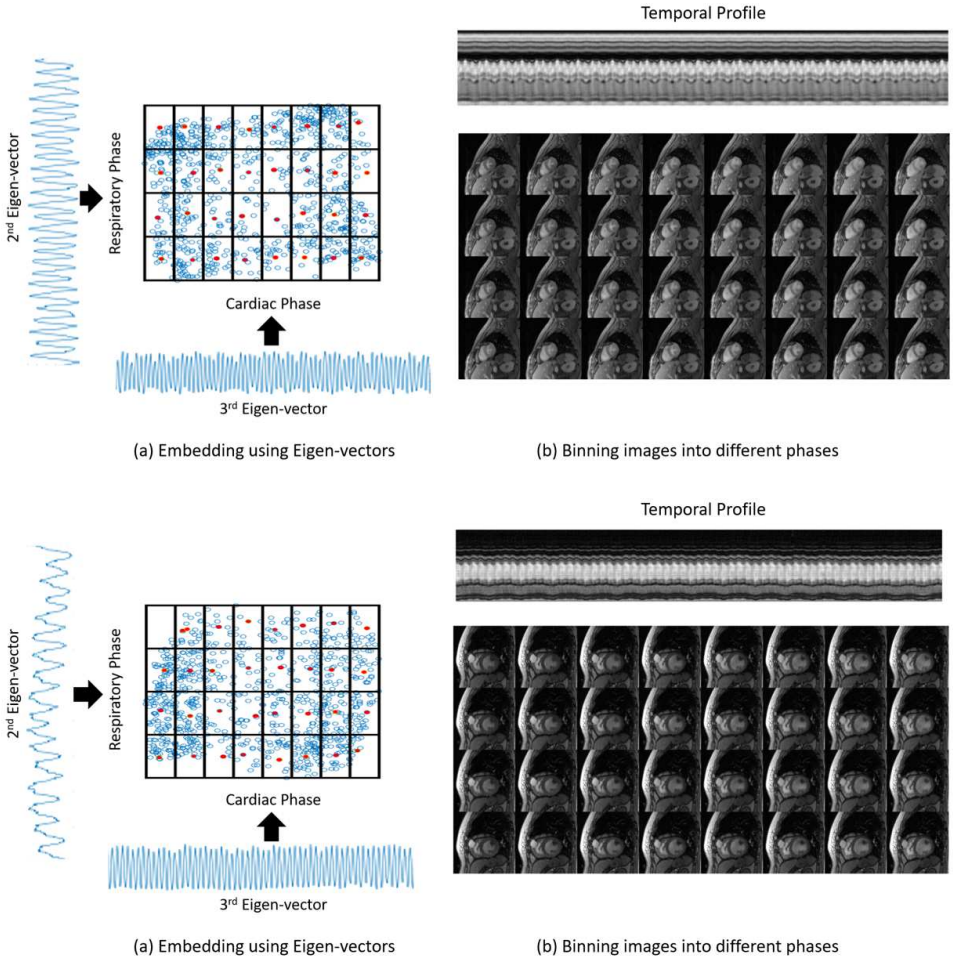


Figure 4.8: Binning into cardiac and respiratory phases. We demonstrate that the reconstructed ungated image series can easily be converted to a gated series of images if desired. For this purpose, the 2^{nd} and 3^{rd} eigen-vectors of the estimated Laplacian matrix are used as an estimate of the respiratory and cardiac phases respectively. The images can then be separated into the desired number of cardiac and respiratory bins. Here, we demonstrate this on two datasets that have been separated into 8 cardiac and 4 respiratory phases. Representative images from these bins have been shown in the figure.

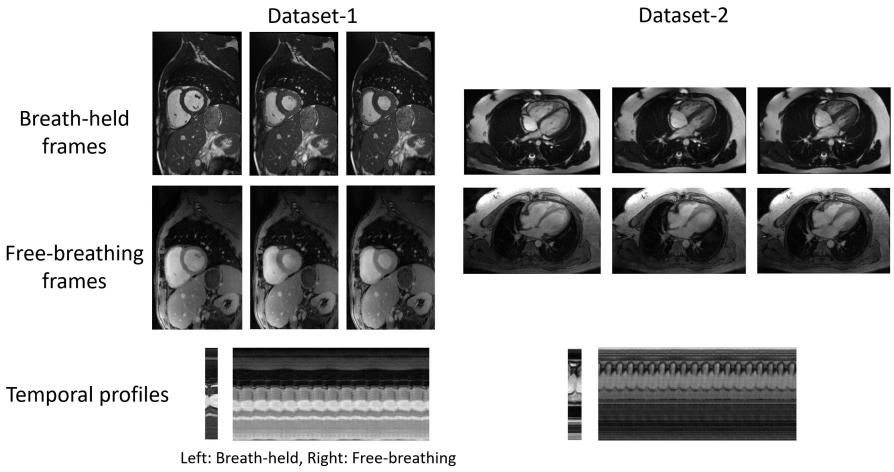


Figure 4.9: Comparison to breath-held scheme. We demonstrate that b-SToRM produces images of similar quality to clinical breath-held scans, in the same acquisition time. Note that there are differences between the free-breathing and breath-held images due to variations in contrast between TRUFI and FLASH acquisitions, and also due to mismatch in slice position. However, the images we obtain are of clinically acceptable quality. Moreover, unlike the breath-held scheme we reconstruct the whole image time series (as is evident from the temporal profile). This can provide richer information, such as studying the interplay of cardiac and respiratory motion.

and gated techniques. There is some difference in the appearance of the breath-held and free-breathing reconstructions due to mismatch in slice position. Moreover, the breath-held acquisition was done using a TRUFI sequence, and thus shows higher contrast than the free-breathing data which was acquired using a FLASH sequence. In spite of these differences, we note that the images reconstructed using b-SToRM are of clinically acceptable quality.

4.4 Discussion

The proposed framework for dynamic image reconstruction has similarities to approaches that rely on the factorization of the Casorati matrix [49, 94]. The key difference is the signal model and the approach in which the temporal basis functions are estimated. Moreover, we have shown in Fig 4.2 that unlike other approaches, b-SToRM is able to automatically estimate the respiratory and cardiac signals from

the eigen-vectors of the estimated Laplacian matrix. When the Laplacian matrix is estimated using the exponential function as in SToRM, or using the navigator signals as in the PSF method, a few eigen-vectors do not capture the physiological motion. For SToRM, it is also required to threshold the weight matrix to achieve good reconstruction results. This is equivalent to heuristically forming a regular graph by fixing the node degree. In this case, the eigen-vectors are dependent on the specific thresholding function that is used. The proposed Laplacian estimation technique does not require any manual thresholding and does not constrain the graph to be a regular one. When reconstructing using a fixed number of basis functions ($r = 30$), it is shown that the proposed Laplacian preserves the temporal profiles better than when an exponential weight matrix is used as in SToRM. Moreover, due to the need to reconstruct only a few basis images, b-SToRM is significantly faster than SToRM. It was illustrated in [68] that an exponential weight matrix can also be used to estimate the respiratory and cardiac signals. However, this was shown for phantom data, and we have found that it does not hold true in general for many real datasets. Other conventional methods often require the binning of the k-space data to respiratory bins before reconstructions, using self gating approaches [25]. The main benefit of b-SToRM is that it does not require any explicit binning. However, we have shown in Fig 4.4 and Fig 4.8 that our reconstructions can easily be visualized in a binned fashion. In contrast, as shown in Fig 4.4, the temporal profiles obtained by rearranging the XD-GRASP reconstructed images often have artefacts due to binning into discrete phases. Thus, when images are reconstructed using a binned approach, they might not always be rearranged to get back the original time series.

We demonstrate our algorithm on a number of datasets with different respiratory and cardiac patterns. In accordance with the results of our retrospective experiments on the impact of the number of navigator lines, we plan to collect data with only two navigator lines in the future. This would increase the incoherence of the un-

dersampling patterns across frames, resulting in better quality reconstructions. Our experiments on reduced scan time show that we can obtain reliable data from datasets with high motion with around 22s of data/slice, while it can be pushed down to 12s for datasets with less motion.

Our method produces a series of ungated images, enabling the user to visualize the real-time data with both respiratory and cardiac motion. This approach may be useful in studies on patients with pulmonary complications such as COPD. The data can also be automatically segmented into respiratory and cardiac phases post reconstruction for easy visualization, using the eigen-vectors of the estimated Laplacian matrix.

Since the study was an add on to the routine cardiac exam, there was no perfect control on the specific time point of acquisition following contrast administration. This explains the differing contrast between the datasets.

4.5 Conclusion

We introduce an algorithm to reconstruct free-breathing and ungated cardiac MR images using a kernel low-rank regularized optimization problem. The success of the method on very challenging datasets with high cardiac rate and irregular breathing patterns suggests a useful clinical application of the method on patients who have difficulty in following traditional breath-holding instructions. It is demonstrated that the resulting ungated images can be easily binned into respiratory and cardiac phases and viewed as a gated dataset. This method shows improved performance and reduced computational complexity over the SToRM algorithm introduced in Chapter 2.

CHAPTER 5
RECOVERY OF POINTS IN CLUSTERS USING FUSION PENALTIES

5.1 Introduction

Clustering is an exploratory data analysis technique that is widely used to discover natural groupings in large datasets, where no labeled or pre-classified samples are available apriori. Specifically, it assigns an object to a group if it is similar to other objects within the group, while being dissimilar to objects in other groups. Example applications include analysis of gene experssion data, image segmentation, identification of lexemes in handwritten text, search result grouping and recommender systems [77]. A wide variety of clustering methods have been introduced over the years; see [37] for a review of classical methods. However, there is no consensus on a particular clustering technique that works well for all tasks, and there are pros and cons to most existing algorithms. The common clustering techniques such as k-means [52], k-medians [8] and spectral clustering [58] are implemented using the Lloyd's algorithm which is non-convex and thus sensitive to initialization. Recently, linear programming and semi-definite programming based convex relaxations of the k-means and k-medians algorithms have been introduced [1] to overcome the dependence on initialization. Unlike the Lloyd's algorithm, these relaxations can provide a certificate of optimality. However, all of the above mentioned techniques require apriori knowledge of the desired number of clusters. Hierarchical clustering methods [92], which produce easily interpretable and visualizable clustering results for a varying number of clusters, have been introduced to overcome the above challenge. A drawback of [92] is its sensitivity to initial guess and perturbations in the data. The more recent convex clustering technique (also known as sum-of-norms clustering) [34] retains the advantages of hierarchical clustering, while being invariant to initialization, and producing a unique clustering path. Theoretical guarantees for successful clustering using the convex-clustering technique are also available [95].

Most of the above clustering algorithms cannot be directly applied to real-life datasets, where a large fraction of samples are missing. For example, gene expression data often contains missing entries due to image corruption, fabrication errors or contaminants [19], rendering gene cluster analysis difficult. Likewise, large databases used by recommender systems (e.g Netflix) usually have a huge amount of missing data, which makes pattern discovery challenging [6]. The presence of missing responses in surveys [9] and failing imaging sensors in astronomy [90] are reported to make the analysis in these applications challenging. Several approaches were introduced to extend clustering to missing-data applications. For example, a partially observed dataset can be converted to a fully observed one using either deletion or imputation [20]. Deletion involves removal of variables with missing entries, while imputation tries to estimate the missing values and then performs clustering on the completed dataset. An extension of the weighted sum-of-norms algorithm (originally introduced for fully sampled data [34]) has been proposed where the weights are estimated from the data points by using some imputation techniques on the missing entries [13]. Kernel-based methods for clustering have also been extended to deal with missing entries by replacing Euclidean distances with partial distances [33,76]. A majorize minimize algorithm was introduced to solve for the cluster-centres and cluster memberships in [16], which offers proven reduction in cost with iteration. In [36] and [47] the data points are assumed to lie on a mixture of K distributions, where K is known. The algorithms then alternate between the maximum likelihood estimation of the distribution parameters and the missing entries. A challenge with these algorithms is the lack of theoretical guarantees for successful clustering in the presence of missing entries. In contrast, there has been a lot of work in recent years on matrix completion for different data models. Algorithms along with theoretical guarantees have been proposed for low-rank matrix completion [11] and subspace clustering from data with missing entries [24], [23]. However, these algorithms and their theoretical

guarantees cannot be trivially extended to the problem of clustering in the presence of missing entries.

The main focus of this work is to introduce an algorithm for the clustering of data with missing entries and to theoretically analyze the conditions needed for perfect clustering in the presence of missing data. The proposed algorithm is inspired by the sum-of-norms clustering technique [34]; it is formulated as an optimization problem, where an auxiliary variable assigned to each data point is an estimate of the centre of the cluster to which that point belongs. A fusion penalty is used to enforce equality between many of these auxiliary variables. Since we have experimentally observed that non-convex fusion penalties provide superior clustering performance, we focus on the analysis of clustering using a ℓ_0 fusion penalty in the presence of missing entries, for an arbitrary number of clusters. The analysis reveals that perfect clustering is guaranteed with high probability, provided the number of measured entries (probability of sampling) is high enough; the required number of measured entries depends on several parameters including intra-cluster variance and inter-cluster distance. We observe that the required number of entries is critically dependent on coherence, which is a measure of the concentration of inter cluster differences in the feature space. Specifically, if the clustering of the points is determined only by a very small subset of all the available features, then the clustering becomes quite unstable if those particular feature values are unknown for some points. Other factors which influence the clustering technique are the number of features, number of clusters and total number of points. We also extend the theoretical analysis to the case without missing entries. The analysis in this setting shows improved bounds when a uniform random distribution of points in their respective clusters is considered, compared to the worst case analysis considered in the missing-data setting. We expect that improved bounds can also be derived for the case with missing data when a uniform random distribution is considered.

We also propose an algorithm to solve a relaxation of the above ℓ_0 penalty based clustering problem, using non-convex saturating fusion penalties. The algorithm is demonstrated on a simulated dataset with different fractions of missing entries and cluster separations. We observe that the algorithm is stable with changing fractions of missing entries, and the clustering performance degrades gradually with an increase in the number of missing entries. We also demonstrate the algorithm on clustering of the Wine dataset [46] and an Australian Sign Language (ASL) dataset [40].

5.2 Clustering using ℓ_0 fusion penalty

5.2.1 Background

We consider the clustering of points drawn from one of K distinct clusters $C_1, C_2, \ldots, C_K$. We denote the center of the clusters by $\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_K \in \mathbb{R}^P$. For simplicity, we assume that there are M points in each of the clusters. The individual points in the k^{th} cluster are modelled as:

$$\mathbf{z}_k(m) = \mathbf{c}_k + \mathbf{n}_k(m); \quad m = 1, .., M, \ k = 1, \ldots, K \tag{5.1}$$

Here, $\mathbf{n}_k(m)$ is the noise or the variation of $\mathbf{z}_k(m)$ from the cluster center $\mathbf{c}_k$. The set of input points $\{\mathbf{x}_i\}, i = 1, .., KM$ is obtained as a random permutation of the points $\{\mathbf{z}_k(m)\}$. The objective of a clustering algorithm is to estimate the cluster labels, denoted by $\mathcal{C}(\mathbf{x}_i)$ for $i = 1, .., KM$.

The sum-of-norms (SON) method is a recently proposed convex clustering algorithm [34]. Here, a surrogate variable $\mathbf{u}_i$ is introduced for each point $\mathbf{x}_i$, which is an estimate of the centre of the cluster to which $\mathbf{x}_i$ belongs. As an example, let $K = 2$ and $M = 5$. Without loss of generality, let us assume that $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_5$ belong to $\mathcal{C}_1$ and $\mathbf{x}_6, \mathbf{x}_7, \ldots, \mathbf{x}_{10}$ belong to $\mathcal{C}_2$. Then, we expect to arrive at the solution: $\mathbf{u}_1 = \mathbf{u}_2 = \ldots = \mathbf{u}_5 = \mathbf{c}_1$ and $\mathbf{u}_6 = \mathbf{u}_7 = \ldots = \mathbf{u}_{10} = \mathbf{c}_2$. In order to find the optimal $\{\mathbf{u}_i^*\}$, the following optimization problem is solved:

$$\{\mathbf{u}_i^*\} = \arg\min_{\{\mathbf{u}_i\}} \sum_{i=1}^{KM} \|\mathbf{x}_i - \mathbf{u}_i\|_2^2 + \lambda \sum_{i=1}^{KM} \sum_{j=1}^{KM} \|\mathbf{u}_i - \mathbf{u}_j\|_p \tag{5.2}$$

The fusion penalty ($\|\mathbf{u}_i - \mathbf{u}_j\|_p$) can be enforced using different ℓ_p norms, out of which the ℓ_1, ℓ_2 and ℓ_∞ norms have been used in literature [34]. The use of sparsity promoting fusion penalties encourages sparse differences $\mathbf{u}_i - \mathbf{u}_j$, which facilitates the clustering of the points $\{\mathbf{u}_i\}$. For an appropriately chosen λ, the $\mathbf{u}_i$'s corresponding to $\mathbf{x}_i$'s from the same cluster converge to the same point. The main benefit of this convex scheme over classical clustering algorithms is the convergence of the algorithm to the global minimum.

The above optimization problem can be solved efficiently using the Alternating Direction Method of Multipliers (ADMM) algorithm and the Alternating Minimization Algorithm (AMA) [15]. Truncated ℓ_1 and ℓ_2 norms have also been used recently in the fusion penalty, resulting in non-convex optimization problems [66]. It has been shown that these penalties provide superior performance to the traditional convex penalties. Convergence to local minimum using an iterative algorithm has also been guaranteed in the non-convex setting.

The sum-of-norms algorithm has also been used as a visualization and exploratory tool to discover patterns in datasets [13]. Clusterpath diagrams are a common way to visualize the data. This involves plotting the solution path as a function of the regularization parameter λ. For a very small value of λ, the solution is given by: $\mathbf{u}_i^* = \mathbf{x}_i$, i.e. each point forms its individual cluster. For a very large value of λ, the solution is given by: $\mathbf{u}_i^* = c$, i.e. every point belongs to the same cluster. For intermediate values of λ, more interesting behaviour is seen as various $\{\mathbf{u}_i\}$ merge and reveal the cluster structure of the data.

In this work, we extend the algorithm to account for missing entries in the data. We present theoretical guarantees for clustering with and without missing entries using an ℓ_0 fusion penalty. Next, we approximate the ℓ_0 penalty by non-convex

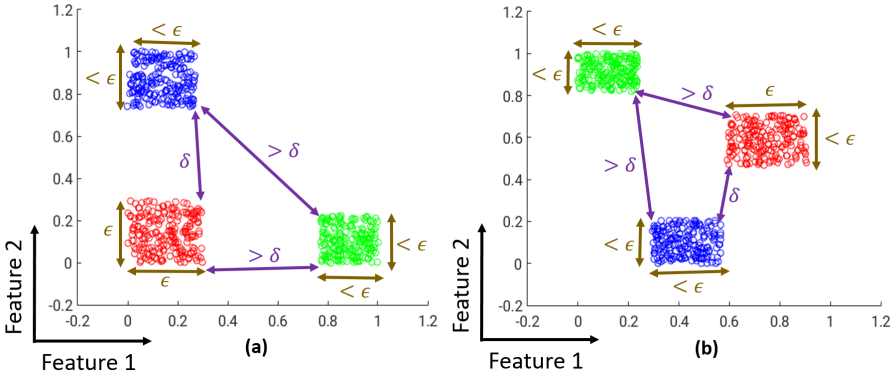


Figure 5.1: Central Assumptions: (a) and (b) illustrate different instances where points belonging to $\mathbb{R}^2$ are to be separated into 3 different clusters (denoted by the colours red, green and blue). Assumptions A.1 and A.2 related to cluster separation and cluster size respectively, are illustrated in both (a) and (b). The importance of assumption A.3 related to feature concentration can also be appreciated by comparing (a) and (b). In (a), points in the red and blue clusters cannot be distinguished solely on the basis of feature 1, while the red and green clusters cannot be distinguished solely on the basis of feature 2. Thus, it is difficult to correctly cluster these points if either of the feature values is unknown. In (b), due to low coherence (as assumed in A.3), this problem does not arise.

saturating penalties, and solve the resulting relaxed optimization problem using an iterative reweighted least squares (IRLS) strategy [12]. The proposed algorithm is shown to perform clustering correctly in the presence of large fractions of missing entries.

5.2.2 Central assumptions

We make the following assumptions (illustrated in Fig 5.1), which are key to the successful clustering of the points:

A.1: Cluster separation: Points from different clusters are separated by $\delta > 0$ in the ℓ_2 sense, i.e:

$$\min_{\{m,n\}} \|\mathbf{z}_k(m) - \mathbf{z}_l(n)\|_2 \geq \delta; \ \forall\ k \neq l \tag{5.3}$$

We require $\delta > 0$ for the clusters to be non-overlapping. A high δ corresponds

to well separated clusters.

A.2: Cluster size: The maximum separation of points within any cluster in the ℓ_∞ sense is $\epsilon \geq 0$, i.e:

$$\max_{\{m,n\}} \|\mathbf{z}_k(m) - \mathbf{z}_k(n)\|_\infty = \epsilon;\ \forall k = 1, \ldots, K \tag{5.4}$$

Thus, the k^{th} cluster is contained within a cube of size ϵ, with center $\mathbf{c}_k$.

A.3: Feature concentration: The coherence of a vector $\mathbf{y} \in \mathbb{R}^P$ is defined as [11]:

$$\mu(\mathbf{y}) = \frac{P\|\mathbf{y}\|_\infty^2}{\|\mathbf{y}\|_2^2} \tag{5.5}$$

By definition: $1 \leq \mu(\mathbf{y}) \leq P$. Intuitively, a vector with a high coherence has a few large values and several small ones. Specifically, if $\mu(\mathbf{y}) = P$, then $\mathbf{y}$ has only 1 non-zero value. In contrast, if $\mu(\mathbf{y}) = 1$, then all the entries of $\mathbf{y}$ are equal. We bound the coherence of the difference between points from different clusters as:

$$\max_{\{m,n\}} \mu(\mathbf{z}_k(m) - \mathbf{z}_l(n)) \leq \mu_0;\ \forall\ k \neq l \tag{5.6}$$

μ_0 is indicative of the difficulty of the clustering problem in the presence of missing data. If $\mu_0 = P$, then two clusters differ only a single feature, suggesting that it is difficult to assign the correct cluster to a point if this feature is not sampled. The best case scenario is $\mu_0 = 1$, when all the features are equally important. In general, cluster recovery from missing data becomes challenging with increasing μ_0.

The quantity $\kappa = \frac{\epsilon\sqrt{P}}{\delta}$ is a measure of the difficulty of the clustering problem. Small values of κ suggest large inter-cluster separation compared to the cluster size; the recovery of such well-defined clusters is expected to be easier than the case with large

κ values. Note the ℓ_2 norm is used in the definition of δ, while the ℓ_∞ norm is used to define ϵ. If $\delta = \epsilon\sqrt{P}$, then $\kappa = 1$; this value of κ is of special importance since $\kappa < 1$ is a requirement for successful recovery in our main results.

We study the problem of clustering the points $\{\mathbf{x}_i\}$ in the presence of entries missing uniformly at random. We arrange the points $\{\mathbf{x}_i\}$ as columns of a matrix $\mathbf{X}$. The rows of the matrix are referred to as features. We assume that each entry of $\mathbf{X}$ is observed with probability p_0. The entries measured in the i^{th} column are denoted by:

$$\mathbf{y}_i = \mathbf{S}_i \, \mathbf{x}_i, \quad i = 1, .., KM \tag{5.7}$$

where $\mathbf{S}_i$ is the sampling matrix, formed by selecting rows of the identity matrix. We consider solving the following optimization problem to obtain the cluster memberships from data with missing entries:

$$\begin{aligned} \{\mathbf{u}_i^*\} = & \min_{\{\mathbf{u}_i\}} \sum_{i=1}^{KM} \sum_{j=1}^{KM} \|\mathbf{u}_i - \mathbf{u}_j\|_{2,0} \\ & \text{s.t } \|\mathbf{S}_i \, (\mathbf{x}_i - \mathbf{u}_i)\|_\infty \leq \frac{\epsilon}{2}, i \in \{1 \dots KM\} \end{aligned} \tag{5.8}$$

We use the above constrained formulation rather than the unconstrained formulation in (5.2) to avoid the dependence on λ. The $\ell_{2,0}$ norm is defined as:

$$\|\mathbf{x}\|_{2,0} = \begin{cases} 0 & \text{, if } \|\mathbf{x}\|_2 = 0 \\ 1 & \text{, otherwise} \end{cases} \tag{5.9}$$

Similar to the SON scheme (5.2), we expect that all $\mathbf{u}_i$'s that correspond to $\mathbf{x}_i$ in the same cluster are equal, while $\mathbf{u}_i$'s from different clusters are not equal. We consider the cluster recovery to be successful when there are no mis-classifications. We claim that the above algorithm can successfully recover the clusters with high probability when:

1. The clusters are well separated (i.e, low $\kappa = \frac{\epsilon\sqrt{P}}{\delta}$)).

2. The sampling probability p_0 is sufficiently high.

3. The coherence μ_0 is small.

Before moving on to a formal statement and proof of this result, we consider a simple special case to illustrate the approach. In order to aid the reader in following the results, all the important symbols used in this work have been summarized in Table 5.1.

5.2.3 Noiseless clusters with missing entries

We consider the simple case where all the points belonging to the same cluster are identical. Thus every cluster is "noiseless", and we have: $\epsilon = 0$ and hence $\kappa = 0$. The optimization problem (5.8) now reduces to:

$$\begin{aligned} \{\mathbf{u}_i^*\} = &\min_{\{\mathbf{u}_i\}} \sum_{i=1}^{KM} \sum_{j=1}^{KM} \|\mathbf{u}_i - \mathbf{u}_j\|_{2,0} \\ &\text{s.t } \mathbf{S}_i \, \mathbf{x}_i = \mathbf{S}_i \, \mathbf{u}_i, i \in \{1 \dots KM\} \end{aligned} \tag{5.10}$$

Next, we state a few results for this special case in order to provide some intuition about the problem. The results are not stated with mathematical rigour and are not accompanied by proofs. In the next sub-section, when we consider the general case, we will provide lemmas and theorems (with proofs in Appendix B), which generalize the results stated here. Specifically, Lemmas 5.2.1, 5.2.2, 5.2.3 and Theorem 5.2.4 generalize Results 5.2.1, 5.2.2, 5.2.3 and 5.2.4 respectively.

We will first consider the data consistency constraint in (5.10) and determine possible feasible solutions. We observe that all the points in any specified cluster can share a centre without violating the data consistency constraint:

Table 5.1: Notations used

K	Number of clusters
M	Number of points in each cluster
P	Number of features for each point
$\mathcal{C}_i$	The i^{th} cluster
$\mathbf{c}_i$	Centre of $\mathcal{C}_i$
$\mathbf{z}_i(m)$	m^{th} point in $\mathcal{C}_i$
$\{\mathbf{x}_i\}$	Random permutation of KM points $\{\mathbf{z}_k(m)\}$ for $k \in \{1, 2, \ldots, K\}, m \in \{1, 2, \ldots, M\}$
$\mathbf{S}_i$	Sampling matrix for $\mathbf{x}_i$
$\mathbf{X}$	Matrix formed by arranging $\{\mathbf{x}_i\}$ as columns, such that the i^{th} column is $\mathbf{x}_i$
p_0	Probability of sampling each entry in $\mathbf{X}$
δ	Parameter related to cluster separation defined in (5.3)
ϵ	Parameter related to cluster size defined in (5.4)
κ	Defined as $\kappa = \frac{\epsilon\sqrt{P}}{\delta}$
μ_0	Parameter related to coherence defined in (5.6)
γ_0	Defined in (5.16)
δ_0	Defined in (5.17)
β_0	Defined in (5.18)
η_0	Defined in (5.19)
$\eta_{0,\text{approx}}$	Upper bound for η_0 for the case of 2 clusters, defined in (5.21)
c	Parameter related to cluster centre separation defined in (5.27)
κ'	Defined as $\kappa' = \frac{\epsilon\sqrt{P}}{c}$
β_1	Defined in (5.28)
η_1	Probability of failure of Theorem 5.2.7

Result 5.2.1. *Consider any two points* $\mathbf{x}_1$ *and* $\mathbf{x}_2$ *from the same cluster. A solution* $\mathbf{u}$ *exists for the following equations:*

$$\mathbf{S}_i \, \mathbf{x}_i \;\; = \;\; \mathbf{S}_i \, \mathbf{u}; \;\; i = 1, 2 \tag{5.11}$$

with probability 1*.*

The proof for the above result is trivial in this special case, since all points in the same cluster are the same. We now consider two points from different clusters.

Result 5.2.2. *Consider two points* $\mathbf{x}_1$ *and* $\mathbf{x}_2$ *from different clusters. A solution* $\mathbf{u}$ *exists for the following equations:*

$$\mathbf{S}_i \, \mathbf{x}_i = \mathbf{S}_i \, \mathbf{u}; \;\; i = 1, 2 \tag{5.12}$$

with low probability, when the sampling probability p_0 *is high and coherence* μ_0 *is low.*

By definition, $\mathbf{S}_1 = \mathbf{S}_{\mathcal{I}_1}$ and $\mathbf{S}_2 = \mathbf{S}_{\mathcal{I}_2}$, where $\mathcal{I}_1$ and $\mathcal{I}_2$ are the index sets of the features that are sampled (not missing) in $\mathbf{x}_1$ and $\mathbf{x}_2$ respectively. We observe that (5.12) can be satisfied, iff:

$$\mathbf{S}_{\mathcal{I}_1 \cap \mathcal{I}_2}(\mathbf{x}_1 - \mathbf{x}_2) = \mathbf{0} \tag{5.13}$$

which implies that the features of $\mathbf{x}_1$ and $\mathbf{x}_2$ are the same on the index set $\mathcal{I}_1 \cap \mathcal{I}_2$. If the probability of sampling p_0 is sufficiently high, then the number of samples at commonly observed locations:

$$|\mathcal{I}_1 \cap \mathcal{I}_2| = q \tag{5.14}$$

will be high, with high probability. If the coherence μ_0 defined in assumption A3 is low, then with high probability the vector $\mathbf{x}_1 - \mathbf{x}_2$ does not have q entries that are equal to 0. In other words, the cluster memberships are not determined by only a few features. Thus, for a small value of μ_0 and high p_0, we can ensure that (5.13)

occurs with very low probability. We now generalize the above result to obtain the following:

Result 5.2.3. *Assume that* $\{\mathbf{x}_i : i \in \mathcal{I}, |\mathcal{I}| = M\}$ *is a set of points chosen randomly from multiple clusters (not all are from the same cluster). A solution* $\mathbf{u}$ *exists for the following equations:*

$$\mathbf{S}_i \, \mathbf{x}_i = \mathbf{S}_i \, \mathbf{u}; \ \forall i \in \mathcal{I} \tag{5.15}$$

with low probability, when the sampling probability p_0 *is high and coherence* μ_0 *is low.*

The key message of the above result is that large mis-classified clusters are highly unlikely. We will show that all feasible solutions containing small mis-classified clusters are associated with higher cost than the correct solution. Thus, we can conclude that the algorithm recovers the ground truth solution with high probability, as summarized by the following result.

Result 5.2.4. *The optimization problem* (5.10) *results in the ground-truth clustering with a high probability if the sampling probability* p_0 *is high and the coherence* μ_0 *is low.*

5.2.4 Noisy clusters with missing entries

We will now consider the general case of noisy clusters with missing entries, and will determine the conditions required for (5.8) to yield successful recovery of clusters. The reasoning behind the proof in the general case is similar to that for the special case discussed in the previous sub-section. Before proceeding to the statement of the lemmas and theorems, we define the following quantities:

- Upper bound for probability that two points have less than $\frac{p_0^2 P}{2}$ commonly observed locations:

$$\gamma_0 \coloneqq (\frac{e}{2})^{-\frac{p_0^2 P}{2}} \tag{5.16}$$

- Given that two points from different clusters have more than $\frac{p_0^2 P}{2}$ commonly observed locations, upper bound for probability that they can yield the same $\mathbf{u}$ without violating the constraints in (5.8):

$$\delta_0 \coloneqq e^{-\frac{p_0^2 P(1-\kappa^2)^2}{\mu_0^2}} \tag{5.17}$$

- Upper bound for probability that two points from different clusters can yield the same $\mathbf{u}$ without violating the constraints in (5.8):

$$\beta_0 \coloneqq 1 - (1-\delta_0)(1-\gamma_0) \tag{5.18}$$

- Upper bound for failure probability of (5.8):

$$\eta_0 \coloneqq \sum_{\{m_j\} \in \mathcal{S}} \left[\beta_0^{\frac{1}{2}(M^2 - \sum_j m_j^2)} \prod_j \binom{M}{m_j} \right] \tag{5.19}$$

where $\mathcal{S}$ is the set of all sets of positive integers $\{m_j\}$ such that: $2 \leq \mathcal{U}(\{m_j\}) \leq K$ and $\sum_j m_j = M$. Here, the function $\mathcal{U}$ counts the number of non-zero elements in a set. For example, if $K = 2$ then $\mathcal{S}$ contains all sets of 2 positive integers $\{m_1, m_2\}$, such that $m_1 + m_2 = M$. Thus, $\mathcal{S} = \{\{1, M-1\}, \{2, M-2\}, \{3, M-3\}, \ldots, \{M-1, 1\}\}$ and (5.19) reduces to:

$$\eta_0 = \sum_{i=1}^{M-1} \left[\beta_0^{i(M-i)} \binom{M}{i}^2 \right] \tag{5.20}$$

- We note that the expression for η_0 is quite involved. Hence, to provide some intuition, we simplify this expression for the special case where there are only two clusters. Under the assumption that $\log \beta_0 \leq \frac{1}{M-1} + \frac{2}{M-2} \log \frac{1}{M-1}$, it can

be shown that η_0 is upper-bounded as:

$$\begin{aligned}\eta_0 &= \sum_{i=1}^{M-1}\left[\beta_0^{i(M-i)}\binom{M}{i}^2\right] \\ &\leq M^3\beta_0^{M-1} \\ &\coloneqq \eta_{0,\text{approx}}\end{aligned} \tag{5.21}$$

The above upper bound is derived in Appendix B.6.

We now state the results for clustering with missing entries in the general noisy case. The following two lemmas are generalizations of Results 5.2.1 and 5.2.2 to the noisy case.

Lemma 5.2.1. *Consider any two points* $\mathbf{x}_1$ *and* $\mathbf{x}_2$ *from the same cluster. A solution* $\mathbf{u}$ *exists for the following equations:*

$$\|\mathbf{S}_i(\mathbf{x}_i - \mathbf{u})\|_\infty \;\leq\; \frac{\epsilon}{2}; \quad i = 1, 2 \tag{5.22}$$

with probability 1*.*

The proof of this lemma is in Appendix B.1.

Lemma 5.2.2. *Consider any two points* $\mathbf{x}_1$ *and* $\mathbf{x}_2$ *from different clusters, and assume that* $\kappa < 1$*. A solution* $\mathbf{u}$ *exists for the following equations:*

$$\|\mathbf{S}_i(\mathbf{x}_i - \mathbf{u})\|_\infty \;\leq\; \frac{\epsilon}{2}; \quad i = 1, 2 \tag{5.23}$$

with probability less than β_0*.*

The proof of this lemma is in Appendix B.3. We note that β_0 decreases with a decrease in κ. A small ϵ implies less variability within clusters and a large δ implies well-separated clusters, together resulting in a low value of κ. Both these

characteristics are desirable for clustering and result in a low value of β_0. This lemma also demonstrates that the coherence assumption is important in ensuring that the sampled entries are sufficient to distinguish between a pair of points from different clusters. As a result, β_0 decreases with a decrease in the value of μ_0. As expected, we also observe that β_0 decreases with an increase in p_0.

The above result can be generalized to consider a large number of points from multiple clusters. If we choose M points such that not all of them belong to the same cluster, then it can be shown that with high probability, they cannot share the same $\mathbf{u}$ without violating the constraints in (5.8). This idea (a generalization of Result 5.2.3) is expressed in the following lemma:

Lemma 5.2.3. *Assume that* $\{\mathbf{x}_i : i \in \mathcal{I}, |\mathcal{I}| = M\}$ *is a set of points chosen randomly from multiple clusters (not all are from the same cluster). If* $\kappa < 1$*, a solution* $\mathbf{u}$ *does not exist for the following equations:*

$$\|\mathbf{S}_i (\mathbf{x}_i - \mathbf{u})\|_\infty \leq \frac{\epsilon}{2}; \quad \forall i \in \mathcal{I} \tag{5.24}$$

with probability exceeding $1 - \eta_0$*.*

The proof of this lemma is in Appendix B.4. We note here, that for a low value of β_0 and a high value of M (number of points in each cluster), we will arrive at a very low value of η_0. Using Lemmas 5.2.1, 5.2.2 and 5.2.3, we now move on to our main result which is a generalization of Result 5.2.4:

Theorem 5.2.4. *If* $\kappa < 1$*, the solution to the optimization problem* (5.8) *is identical to the ground-truth clustering with probability exceeding* $1 - \eta_0$*.*

The proof of the above theorem is in Appendix B.5. The reasoning follows from Lemma 5.2.3. It is shown in the proof that all solutions with cluster sizes smaller than M are associated with a higher cost than the ground-truth solution.

5.2.5 Clusters without missing entries

We now study the case where there are no missing entries. In this special case, optimization problem (5.8) reduces to:

$$\begin{aligned}\{\mathbf{u}_i^*\} = &\min_{\{\mathbf{u}_i\}} \sum_{i=1}^{KM}\sum_{j=1}^{KM} \|\mathbf{u}_i - \mathbf{u}_j\|_{2,0} \\ &\text{s.t } \|\mathbf{x}_i - \mathbf{u}_i\|_\infty \leq \frac{\epsilon}{2},\ i \in \{1 \dots KM\}\end{aligned} \tag{5.25}$$

We have the following theorem guaranteeing successful recovery for clusters without missing entries:

Theorem 5.2.5. *If $\kappa < 1$, the solution to the optimization problem* (5.25) *is identical to the ground-truth clustering.*

The proof for the above Theorem is in Appendix B.7. We note that the above result does not consider any particular distribution of the points in each cluster. Instead, if we consider that the points in each cluster are sampled from certain particular probability distributions such as the uniform random distribution, then a larger κ is sufficient to ensure success with high probability. In the general case where no such distribution is assumed, we cannot make a probabilistic argument, and a smaller κ is required. We now consider a special case, where the noise $\mathbf{n}_k(m)$ is a zero mean uniform random variable $\sim U(-\epsilon/2, \epsilon/2)$. Thus, the points within each cluster are uniformly distributed in a cube of side ϵ. We note that δ is now a random variable, and thus instead of using the constant $\kappa = \frac{\epsilon\sqrt{P}}{\delta}$ (as in previous lemmas), we define the following constant:

$$\kappa' = \frac{\epsilon\sqrt{P}}{c} \tag{5.26}$$

where c is defined as the minimum separation between the centres of any 2 clusters in the dataset:

$$\min_{\{k,l\}} \|\mathbf{c}_k - \mathbf{c}_l\|_2 \geq c;\ \forall\ k \neq l \tag{5.27}$$

We also define the following quantity:

$$\beta_1 = e^{-\frac{P(1-\frac{5}{6}\kappa'^2)^2}{8\kappa'^2}} \tag{5.28}$$

We arrive at the following result for two points in different clusters:

Lemma 5.2.6. *Let $\kappa' < \sqrt{\frac{6}{5}}$. If the points in each cluster follow a uniform random distribution, then for two points $\mathbf{x}_1$ and $\mathbf{x}_2$ belonging to different clusters, a solution $\mathbf{u}$ exists for the following equations:*

$$\|\mathbf{x}_i - \mathbf{u}\|_\infty \ \leq \ \frac{\epsilon}{2}; \quad i = 1, 2 \tag{5.29}$$

with probability less than β_1.

The proof for the above lemma is in Appendix B.8. This implies that for $\kappa' < \sqrt{\frac{6}{5}}$, two points from different clusters cannot be misclassified to a single cluster with high probability. As η_0 is expressed in terms of β_0 in (5.19), we can also express η_1 in terms of β_1. We get the following guarantee for perfect clustering:

Theorem 5.2.7. *If the points in each cluster follow a uniform random distribution and $\kappa' < \sqrt{\frac{6}{5}}$, then the solution to the optimization problem* (5.25) *is identical to the ground-truth clustering with probability exceeding $1 - \eta_1$.*

Note that $\kappa = \kappa' \frac{c}{\delta}$. Thus, the above result allows for values $\kappa > 1$. Our results show that if we do not consider the distribution of the points, then we arrive at the bound $\kappa < 1$ with and without missing entries, as seen from Theorems 5.2.4 and 5.2.5 respectively. A uniform random distribution can also be assumed in the case of missing entries. Similar to Theorem 5.2.7, we expect an improved bound for the case with missing entries as well.

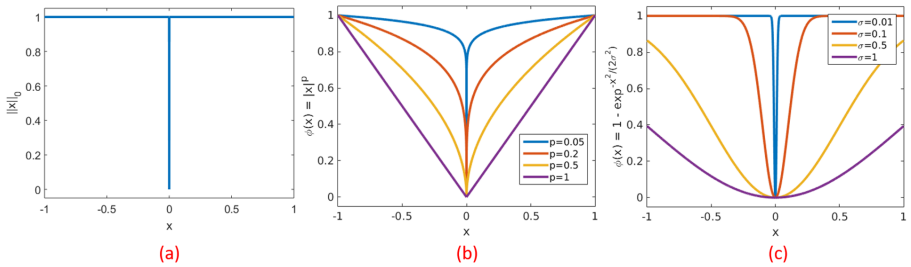


Figure 5.2: Different penalty functions ϕ. (a) The ℓ_0 norm (b) The ℓ_p penalty function which is non-convex for $0 < p < 1$ and convex for $p = 1$ (c) The H_1 penalty function. The ℓ_p and H_1 penalties closely approximate the ℓ_0 norm for low values of p and σ respectively.

5.3 Relaxation of the ℓ_0 penalty

5.3.1 Constrained formulation

We propose to solve a relaxation of the optimization problem (5.8), which is more computationally feasible. The relaxed problem is given by:

$$\begin{aligned} \{\mathbf{u}_i^*\} = &\min_{\{\mathbf{u}_i\}} \sum_{i=1}^{KM} \sum_{j=1}^{KM} \phi\left(\|\mathbf{u}_i - \mathbf{u}_j\|_2\right) \\ &\text{s.t } \|\mathbf{S}_i(\mathbf{x}_i - \mathbf{u}_i)\|_\infty \leq \frac{\epsilon}{2}, i \in \{1 \dots KM\} \end{aligned} \tag{5.30}$$

where ϕ is a function approximating the ℓ_0 norm. Some examples of such functions are:

- ℓ_p norm: $\phi(x) = |x|^p$, for some $0 < p < 1$.
- H_1 penalty: $\phi(x) = 1 - e^{-\frac{x^2}{2\sigma^2}}$.

These functions approximate the ℓ_0 penalty more accurately for lower values of p and σ, as illustrated in Fig 5.2. We reformulate the problem using a majorize-minimize strategy. Specifically, by majorizing the penalty ϕ using a quadratic surrogate functional, we obtain:

$$\phi(x) \leq w(x)x^2 + d \tag{5.31}$$

where $w(x) = \frac{\phi^{'}(x)}{2x}$, and d is a constant. For the two penalties considered here, we obtain the weights as:

- ℓ_p norm: $w(x) = (\frac{2}{p}x^{(2-p)} + \alpha)^{-1}$. The infinitesimally small α term is introduced to deal with situations where $x = 0$. For non-zero x, we get the expression $w(x) \approx \frac{p}{2}x^{p-2}$.
- H_1 penalty: $w(x) = \frac{1}{2\sigma^2}e^{-\frac{x^2}{2\sigma^2}}$.

We can now state the majorize-minimize formulation for problem (5.30) as:

$$\begin{aligned} \{\mathbf{u}_i^*, w_{ij}^*\} = \arg\min_{\{\mathbf{u}_i, w_{ij}\}} & \sum_{i=1}^{KM}\sum_{j=1}^{KM} w_{ij} \ \|\mathbf{u}_i - \mathbf{u}_j\|_2^2 \\ \text{s.t } & \|\mathbf{S}_i(\mathbf{x}_i - \mathbf{u}_i)\|_\infty \leq \ \frac{\epsilon}{2}, i \in \{1 \dots KM\} \end{aligned} \tag{5.32}$$

where the constant d has been ignored. In order to solve problem (5.32), we alternate between two sub-problems till convergence. At the n^{th} iteration, these sub-problems are given by:

$$w_{ij}^{(n)} = \frac{\phi^{'}\left(\|\mathbf{u}_i^{(n-1)} - \mathbf{u}_j^{(n-1)}\|_2\right)}{2\|\mathbf{u}_i^{(n-1)} - \mathbf{u}_j^{(n-1)}\|_2} \tag{5.33}$$

$$\begin{aligned} \{\mathbf{u}_i^{(n)}\} = \arg\min_{\{\mathbf{u}_i\}} & \sum_{i=1}^{KM}\sum_{j=1}^{KM} w_{ij}^{(n)}\|\mathbf{u}_i - \mathbf{u}_j\|_2^2 \\ \text{s.t } & \|\mathbf{S}_i(\mathbf{x}_i - \mathbf{u}_i)\|_\infty \leq \ \frac{\epsilon}{2}, i \in \{1 \dots KM\} \end{aligned} \tag{5.34}$$

5.3.2 Unconstrained formulation

For larger datasets, it might be computationally intensive to solve the constrained problem. In this case, we propose to solve the following unconstrained problem:

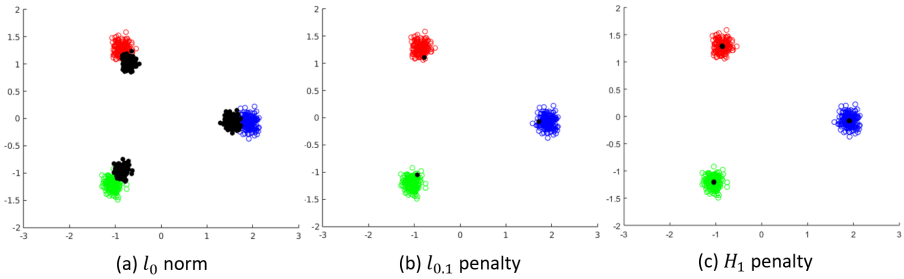


Figure 5.3: Comparison of different penalties. We show here the 2 most significant principal components of the solutions obtained using the IRLS algorithm. (a) It can be seen that the ℓ_1 penalty is unable to cluster the points even though the clusters are well-separated. (b) The $\ell_{0.1}$ penalty is able to cluster the points correctly. However, the cluster-centres are not correctly estimated. (c) The H_1 penalty correctly clusters the points and also gives a good estimate of the centres.

$$\{\mathbf{u}_i^*\} = \arg\min_{\{\mathbf{u}_i\}} \sum_{i=1}^{KM} \|\mathbf{S}_i(\mathbf{u}_i - \mathbf{x}_i)\|_2^2 + \lambda \sum_{i=1}^{KM}\sum_{j=1}^{KM} \phi(\|\mathbf{u}_i - \mathbf{u}_j\|_2) \tag{5.35}$$

As before, we can state the majorize-minimize formulation for problem (5.35) as:

$$\begin{aligned}\{\mathbf{u}_i^*, w_{ij}^*\} = \arg\min_{\{\mathbf{u}_i, w_{ij}\}} &\sum_{i=1}^{KM} \|\mathbf{S}_i(\mathbf{u}_i - \mathbf{x}_i)\|_2^2 \\ &+\lambda \sum_{i=1}^{KM}\sum_{j=1}^{KM} w_{ij}\|\mathbf{u}_i - \mathbf{u}_j\|_2^2\end{aligned} \tag{5.36}$$

In order to solve the problem (5.36), we alternate between two sub-problems till convergence. The 1^{st} sub-problem is the same as (5.33). The 2^{nd} sub-problem is given by:

$$\begin{aligned}\{\mathbf{u}_i^{(n)}\} = \arg\min_{\{\mathbf{u}_i\}} &\sum_{i=1}^{KM} \|\mathbf{S}_i(\mathbf{u}_i - \mathbf{x}_i)\|_2^2 \\ &+\lambda \sum_{i=1}^{KM}\sum_{j=1}^{KM} w_{ij}^{(n)}\|\mathbf{u}_i - \mathbf{u}_j\|_2^2\end{aligned} \tag{5.37}$$

5.3.3 Comparison of penalties

We compare the performance of different penalties when used as a surrogate for the ℓ_0 norm. For this purpose, we use a simulated dataset with points in $\mathbb{R}^{50}$ belonging to 3 well-separated clusters, with 200 points in each cluster. For this particular experiment, we considered $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{200} \in \mathcal{C}_1$, $\mathbf{x}_{201}, \mathbf{x}_{202}, \ldots, \mathbf{x}_{400} \in \mathcal{C}_2$ and $\mathbf{x}_{401}, \mathbf{x}_{402}, \ldots, \mathbf{x}_{600} \in \mathcal{C}_3$. We do not consider the presence of missing entries for this experiment. We solve problem (5.35) to cluster the points using the ℓ_1, ℓ_p (for $p = 0.1$) and H_1 (for $\sigma = 0.5$) penalties. The results are shown in Fig 5.3. Only for the purpose of visualization, we take a PCA of the data matrix $\mathbf{X} \in \mathbb{R}^{50\times 600}$ and retain the 2 most significant principal components to get a matrix of points $\in \mathbb{R}^{2\times 600}$. These points are plotted in the figure, with red, blue and green representing points from different clusters. We similarly obtain the 2 most significant components of the estimated centres and plot the resulting points in black. In (b) and (c), we note that $\mathbf{u}_1^* = \mathbf{u}_2^* = \ldots = \mathbf{u}_{200}^*$, $\mathbf{u}_{201}^* = \mathbf{u}_{202}^* = \ldots = \mathbf{u}_{400}^*$ and $\mathbf{u}_{401}^* = \mathbf{u}_{402}^* = \ldots = \mathbf{u}_{600}^*$. Thus, the ℓ_p penalty and the H_1 penalty are able to correctly cluster the points. This behaviour is not seen in (a). Thus it is concluded that the convex ℓ_1 penalty is unable to cluster the points.

The cluster-centres estimated using the ℓ_p penalty are inaccurate. The H_1 penalty out-performs the other two penalties and accurately estimates the cluster-centres. We can explain this behaviour intuitively by observing the plots in Fig 5.2. The ℓ_1 norm penalizes differences between all pairs of points. The $\ell_{0.1}$ semi-norm penalizes differences between points that are close. Due to the saturating nature of the penalty, it does not heavily penalize differences between points that are further away. The same is true for the H_1 penalty. However, we note that the H_1 penalty saturates to 1 very quickly, similar to the ℓ_0 norm. This behaviour is missing for the $\ell_{0.1}$ penalty. For this reason, it is seen that the $\ell_{0.1}$ penalty also penalizes inter-cluster distances (unlike the H_1 penalty), and shrinks the distance between the estimated centres of

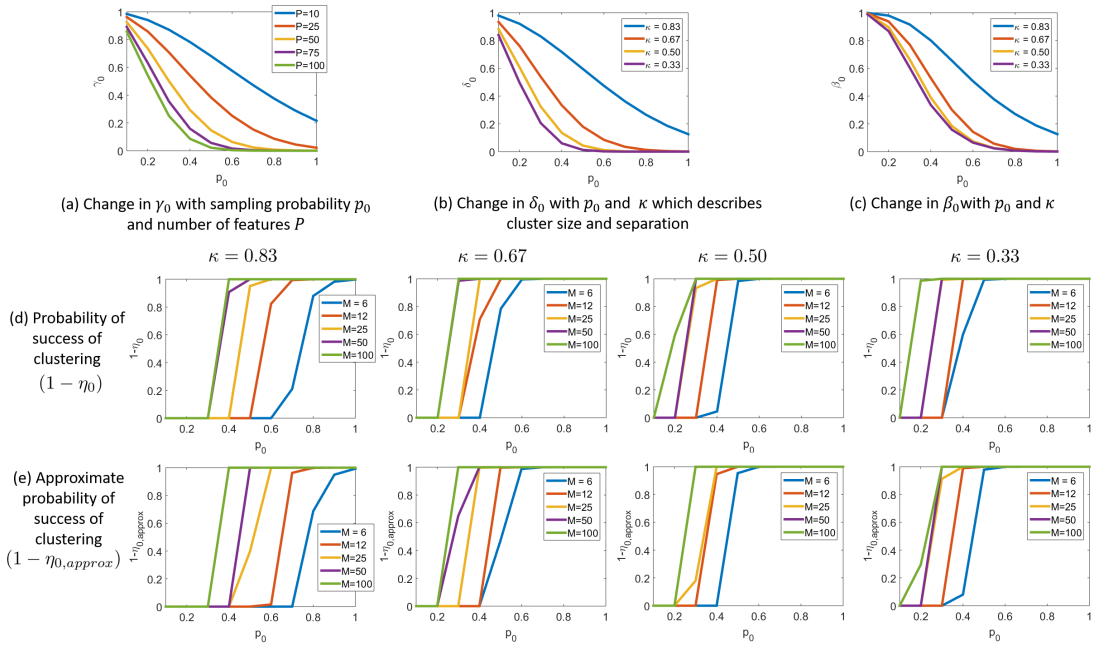


Figure 5.4: Study of Theoretical Guarantees. The quantities γ_0, δ_0 and β_0 defined in Section 5.2.4 are studied in (a), (b) and (c) respectively. In (b) and (c), $P = 50$ and $\mu_0 = 1.5$ are assumed. β_0 gives the probability that 2 points from different clusters can share a centre. As expected, this value decreases with increase in p_0 and decrease in κ. Considering $K = 2$ clusters, a lower bound for the probability of successful clustering $(1 - \eta_0)$ using the proposed algorithm is shown in (d) for different values of κ. The approximate values $(1 - \eta_{0,\text{approx}})$ computed using (5.21) are shown in (e).

different clusters.

5.3.4 Initialization strategies

Our experiments emphasize the need for a good initialization of the weights w_{ij} for convergence to the correct cluster centre estimates. This dependence on the initial value arises from the non-convexity of the optimization problem. We consider two different strategies for initializing the weights:

- Partial Distances: Consider a pair of points $\mathbf{x}_1, \mathbf{x}_2$ observed by sampling matrices $\mathbf{S}_1 = \mathbf{S}_{\mathcal{I}_1}$ and $\mathbf{S}_2 = \mathbf{S}_{\mathcal{I}_2}$ respectively. Let the set of common indices be $\omega := \mathcal{I}_1 \cap \mathcal{I}_2$. We define the partial distance as $\|\mathbf{y}_\omega\| = \sqrt{\frac{P}{|\omega|}}\|\mathbf{x}_{1\omega} - \mathbf{x}_{2\omega}\|$, where

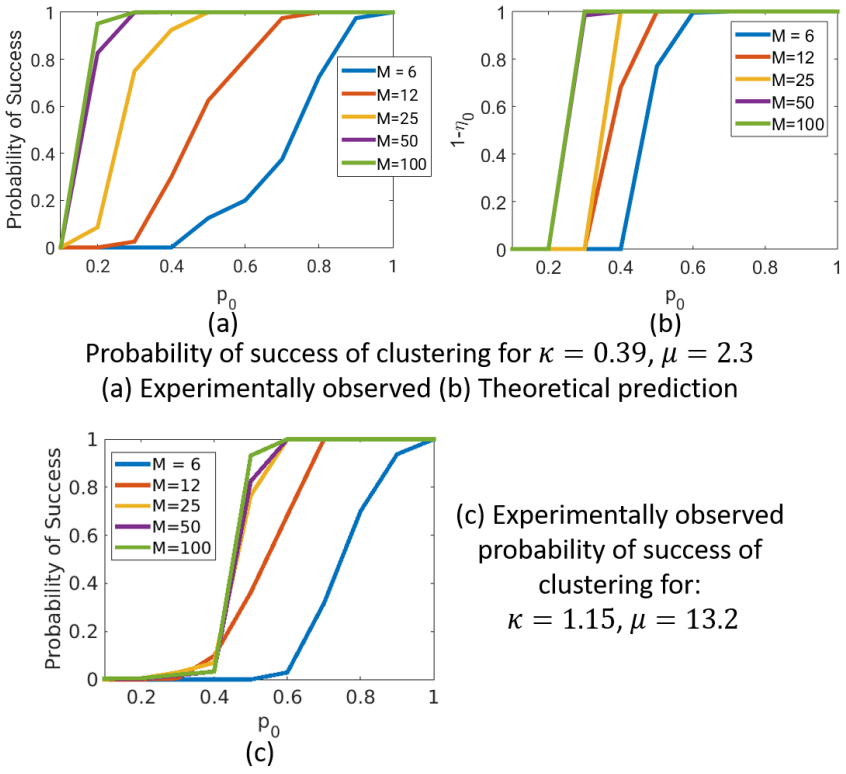


Figure 5.5: Experimental results for probability of success. Guarantees are shown for a simulated dataset with $K = 2$ clusters. The clustering was performed using (5.32) with an H_1 penalty and partial distance based initialization. For (a) and (b) it is assumed that $\kappa = 0.39$ and $\mu_0 = 2.3$. (a) shows the experimentally obtained probability of success of clustering for clusters with points from a uniform random distribution. (b) shows the theoretical lower bound for the probability of success. (c) shows the experimentally obtained probability of success for a more challenging dataset with $\kappa = 1.15$ and $\mu_0 = 13.2$. Note that we do not have theoretical guarantees for this case, since our analysis assumes that $\kappa < 1$.

$\mathbf{x}_{i\omega}$ represents the set of entries of $\mathbf{x}_i$ restricted to the index set ω. Instead of the actual distances which are not available, the partial distances $\|\mathbf{y}_\omega\|$ can be used for computing the weights.

- Imputation Methods: The weights can be computed from estimates $\{\mathbf{u}_i^{(0)}\}$, where:

$$\mathbf{u}_i^{(0)} = \mathbf{S}_i\mathbf{x}_i + (\mathbf{I} - \mathbf{S}_i)\mathbf{m} \tag{5.38}$$

  Here $\mathbf{m}$ is a constant vector, specific to the imputation technique. The zero-filling technique corresponds to $\mathbf{m} = \mathbf{0}$. Better estimation techniques can be derived where the j^{th} row of $\mathbf{m}$ can be set to the mean of all measured values in the j^{th} row of $\mathbf{X}$.

We will observe experimentally that for a good approximation of the initial weights $\mathbf{W}^{(0)}$, we get the correct clustering. Conversely, the clustering fails for a bad initial guess. Our experiments demonstrate the superiority of a partial distance based initialization strategy over a zero-filled initialization.

5.4 Results

We study the proposed theoretical guarantees for Theorem 5.2.4 for different settings. We also test the proposed algorithm on simulated and real datasets. The simulations are used to study the performance of the algorithm with change in parameters such as fraction of missing entries, number of points to be clustered etc. We also study the effect of different initialization techniques on the algorithm performance. We demonstrate the algorithm on the publicly available Wine [46] and ASL [40] datasets.

5.4.1 Study of theoretical guarantees

We observe the behaviour of the quantities $\gamma_0, \delta_0, \beta_0, \eta_0$ and $\eta_{0,\text{approx}}$ (defined in section 5.2.4) as a function of parameters p_0, P, κ and M. Fig 5.4 shows a few plots that illustrate the change in these quantities as the different parameters are varied. γ_0 is an upper bound for the probability that a pair of points have $< \frac{p_0^2 P}{2}$ entries observed at common locations. In Fig 5.4 (a), the change in γ_0 is shown as a function of p_0 for different values of P. In subsequent plots, we fix $P = 50$ and $\mu_0 = 1.5$. δ_0 is an upper bound for the probability that a pair of points from different clusters can share a common centre, given that $\geq \frac{p_0^2 P}{2}$ entries are observed at common locations. In Fig 5.4 (b), the change in δ_0 is shown as a function of p_0 for different values of κ. In Fig 5.4 (c), the behaviour of $\beta_0 = 1 - (1 - \gamma_0)(1 - \delta_0)$ is shown, which is the probability mentioned in Lemma 5.2.2.

We consider the two cluster setting, (i.e. $K = 2$) for subsequent plots. η_0 is the probability of failure of the clustering algorithm (5.8). In (d) and (e), plots are shown for $(1 - \eta_0)$ and $(1 - \eta_{0,\text{approx}})$ as a function of p_0 for different values of κ and M. Here, $\eta_{0,\text{approx}}$ is an upper bound for η_0 computed using (5.21). As expected, the probability of success of the clustering algorithm increases with increase in p_0 and M and decrease in κ.

5.4.2 Clustering of simulated data

We simulated datasets with $K = 2$ disjoint clusters in $\mathbb{R}^{50}$ with a varying number of points per cluster ($M = 6, 12, 25, 50, 100$). The points in each cluster follow a uniform random distribution. We study the probability of success of the H_1 penalty based constrained clustering algorithm (with partial-distance based initialization) as a function of κ, M and p_0. For a particular set of parameters the experiment was conducted 20 times to compute the probability of success of the algorithm. Between these 20 trials, the cluster-centers remain the same, while the points sampled from these clusters are different and the locations of the missing entries are different. Fig

5.5 (a) shows the result for datasets with $\kappa = 0.39$ and $\mu_0 = 2.3$. The theoretical guarantees for successfully clustering the dataset are shown in (b). Note that the theoretical guarantees do not assume that the points are taken from a uniform random distribution. Also, the theoretical bounds assume that we are solving the original problem using a ℓ_0 norm, whereas the experimental results were generated for the H_1 penalty. Our theoretical guarantees hold for $\kappa < 1$. However, we demonstrate in (c) that even for the more challenging case where $\kappa = 1.15$ and $\mu_0 = 13.2$, our clustering algorithm is successful. Note that we do not have theoretical guarantees for this case. However, by assuming a uniform random distribution on the points, we expect that we can get better theoretical guarantees (similar to Theorem 5.2.7 for the case without missing entries).

Clustering results with $K = 3$ simulated clusters are shown in Fig 5.6. We simulated Dataset-1 with $K = 3$ disjoint clusters in $\mathbb{R}^{50}$ and $M = 200$ points in each cluster. In order to generate this dataset, 3 cluster centres in $\mathbb{R}^{50}$ were chosen from a uniform random distribution. The distances between the 3 pairs of cluster-centres are 3.5, 2.8 and 3.3 units respectively. For each of these 3 cluster centres, 200 noisy instances were generated by adding zero-mean white Gaussian noise of variance 0.1. The dataset was sub-sampled with varying fractions of missing entries ($p_0 = 1, 0.9, 0.8, \ldots, 0.3, 0.2$). The locations of the missing entries were chosen uniformly at random from the full data matrix. We also generate Dataset-2 by halving the distance between the cluster centres, while keeping the intra-cluster variance fixed. We test both the constrained (5.30) and unconstrained (5.35) formulations of our algorithm on these datasets. Both the proposed initialization techniques for the IRLS algorithm (i.e. zero-filling and partial-distance) are also tested here. Since the points lie in $\mathbb{R}^{50}$, we take a PCA of the points and their estimated centres (similar to Fig 5.3) and plot the 2 most significant components. The 3 colours distinguish the points according to their ground-truth clusters. Each point $\mathbf{x}_i$ is joined to its

centre estimate $\mathbf{u}_i^*$ by a line. As expected, we observe that the clustering algorithms are more stable with fewer missing entries. We also note that the results are quite sensitive to the initialization technique. We observe that the partial distance based initialization technique out-performs the zero-filled initialization. The unconstrained algorithm with partial distance-based initialization shows superior performance to the alternative schemes. Thus, we use this scheme for subsequent experiments on real datasets.

5.4.3 Clustering of wine dataset

We apply the clustering algorithm to the Wine dataset [46]. The data consists of the results of a chemical analysis of wines from 3 different cultivars. Each data point has $P = 13$ features. The 3 clusters have 59, 71 and 48 points respectively, resulting in a total of 178 data points. We created a dataset without outliers by retaining only $M = 40$ points per cluster, resulting in a total of 120 data points. We under-sampled these datasets using uniform random sampling with different fractions of missing entries. The results are displayed in Fig 5.7 using the PCA technique as explained in the previous sub-section. It is seen that the clustering is quite stable and degrades gradually with increasing fractions of missing entries.

5.4.4 Clustering of ASL dataset

We apply the clustering algorithm to subsets of words from the Australian Sign Language high quality dataset [40]. The original dataset contained 2565 signs, each repeated 27 times by a single user over a period of 9 weeks. 28 features are measured for each sign, with an average length of 57 time frames for each feature. These features correspond to the relative positions and orientations of the fingers, measured using gloves and magnetic position trackers. We picked the most important frame for each frame, resulting in feature vector of length 28 for each word. We next formed two datasets containing subsets of words. The first dataset contained all instances of the

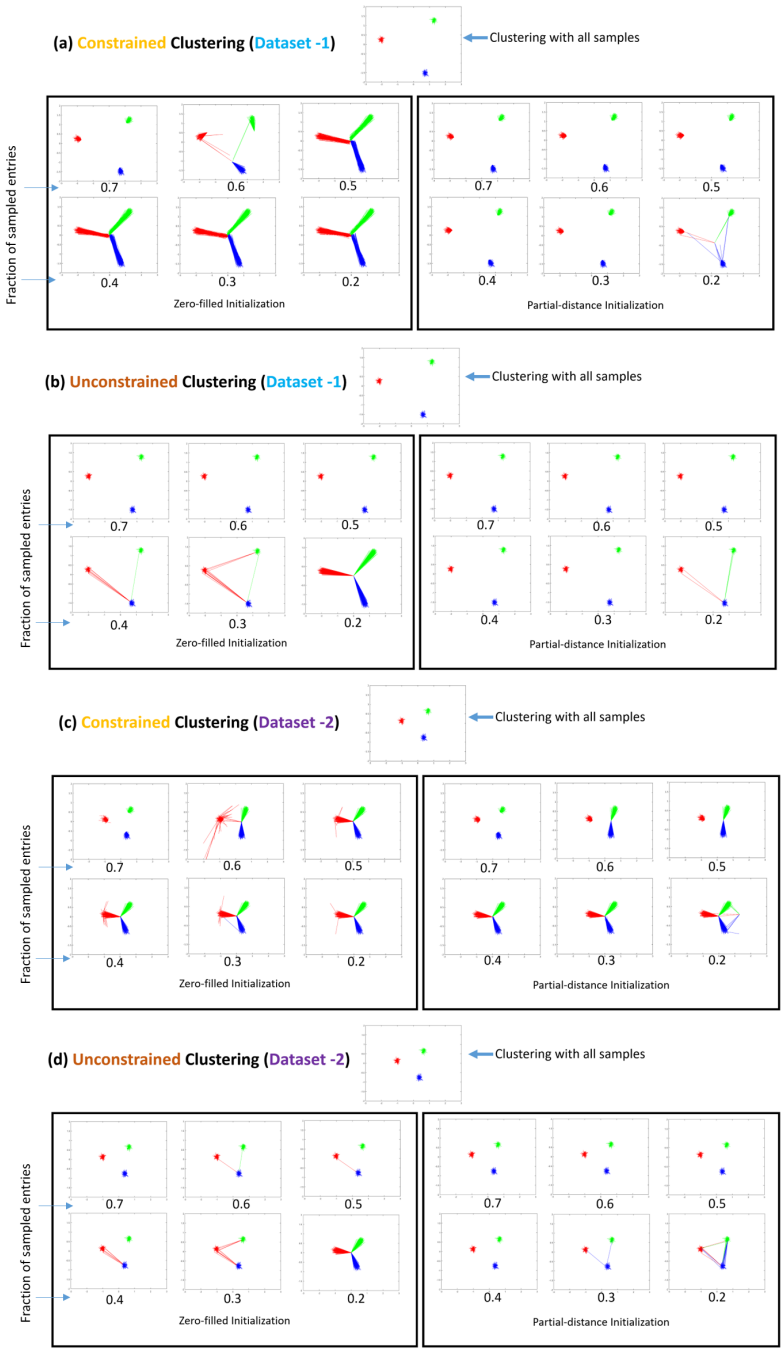


Figure 5.6: Clustering results in simulated datasets. The H_1 penalty is used to cluster two datasets with varying fractions of missing entries. Both the constrained and unconstrained formulation results are presented with different initialization techniques (zero-filled and partial-distance based). We show here the 2 most significant principal components of the solutions. The original points $\{\mathbf{x}_i\}$ are connected to their cluster centre estimates $\{\mathbf{u}_i\}$ by lines. Inter-cluster distances in Dataset 2 are half of those in Dataset 1, while intra-cluster distances remain the same. Consequently, Dataset 1 performs better at a higher fraction of missing entries. For the unconstrained clustering formulation with partial-distance based initialization, the cluster centre estimates are relatively stable with varying fractions of missing entries.

four words "alive", "answer", "boy" and "cold". The second dataset contained all instances of the four words "alive", "boy", "change" and "love". For each dataset, the feature vectors were arranged as columns of the matrix $\mathbf{X}$. Both the datasets were of size 28×108. The datasets were undersampled uniformly at random using different fractions of missing entries. The results are displayed in Fig 5.8 for both datasets. It is observed that clustering the first dataset in the presence of missing entries is relatively easier, since the words more well-separated, as is predicted by theory.

5.5 Discussion

We have proposed a technique to cluster points when some of the feature values of all the points are unknown. We theoretically studied the performance of an algorithm that minimizes an ℓ_0 fusion penalty subject to certain constraints relating to consistency with the known features. We concluded that under favourable clustering conditions, such as well-separated clusters with low intra-cluster variance, the proposed method performs the correct clustering even in the presence of missing entries. However, since the problem is NP-hard, we propose to use other penalties that approximate the ℓ_0 norm. We observe experimentally that the H_1 penalty is a good surrogate for the ℓ_0 norm. This non-convex saturating penalty is shown to perform better in the clustering task than previously used convex norms and penalties. We describe an IRLS based strategy to solve the relaxed problem using the surrogate penalty.

Our theoretical analysis reveals the various factors that determine whether the points will be clustered correctly in the presence of missing entries. It is obvious that the performance degrades with the decrease in the fraction of sampled entries (p_0). Moreover, it is shown that the difference between points from different clusters should have low coherence (μ_0). This means that the expected clustering should not be dependent on only a few features of the points. Intuitively, if the points in different clusters can be distinguished by only 1 or 2 features, then a point missing

these particular feature values cannot be clustered correctly. Moreover, we note that a high number of points per cluster (M), high number of features (P) and a low number of clusters (K) make the data less sensitive to missing entries. Finally, well-separated clusters with low intra-cluster variance (resulting in low values of κ) are desirable for correct clustering.

Our experimental results show great promise for the proposed technique. In particular, for the simulated data, we note that the cluster-centre estimates degrade gradually with increase in the fraction of missing entries. Depending on the characteristics of the data such as number of points and cluster separation distance, the clustering algorithm fails at some particular fraction of missing entries. We also show the importance of a good initialization for the IRLS algorithm, and our proposed initialization technique using partial distances is shown to work very well.

Our theory assumes well-separated clusters and does not consider the presence of any outliers. Theoretical and experimental analysis for the clustering performance in the presence of outliers needs to be investigated. Improving the algorithm performance in the presence of outliers is a direction for future work. Moreover, we have shown improved bounds for the clustering success in the absence of missing entries when the points within a cluster are assumed to follow a uniform random distribution. We expect this trend to also hold for the case with missing entries. This case will be analyzed in future work.

5.6 Conclusion

We propose a clustering technique for data in the presence of missing entries. We prove theoretically that a constrained ℓ_0 norm minimization problem recovers the clustering correctly even in the presence of missing entries. An efficient algorithm that solves a relaxation of the above problem is presented next. It is demonstrated that the cluster centre estimates obtained using the proposed algorithm degrade gradually with an increase in the number of missing entries. The algorithm is also used to

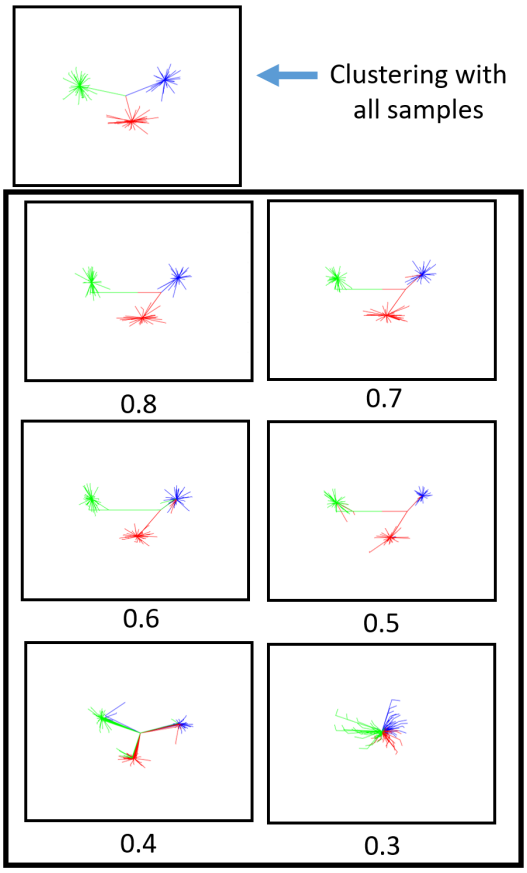


Figure 5.7: Clustering on Wine dataset. The H_1 penalty is used to cluster the Wine datasets with varying fractions of missing entries.

cluster the Wine and ASL datasets. The presented theory and results demonstrate the utility of the proposed algorithm in clustering data when some of the feature values of the data are unknown.

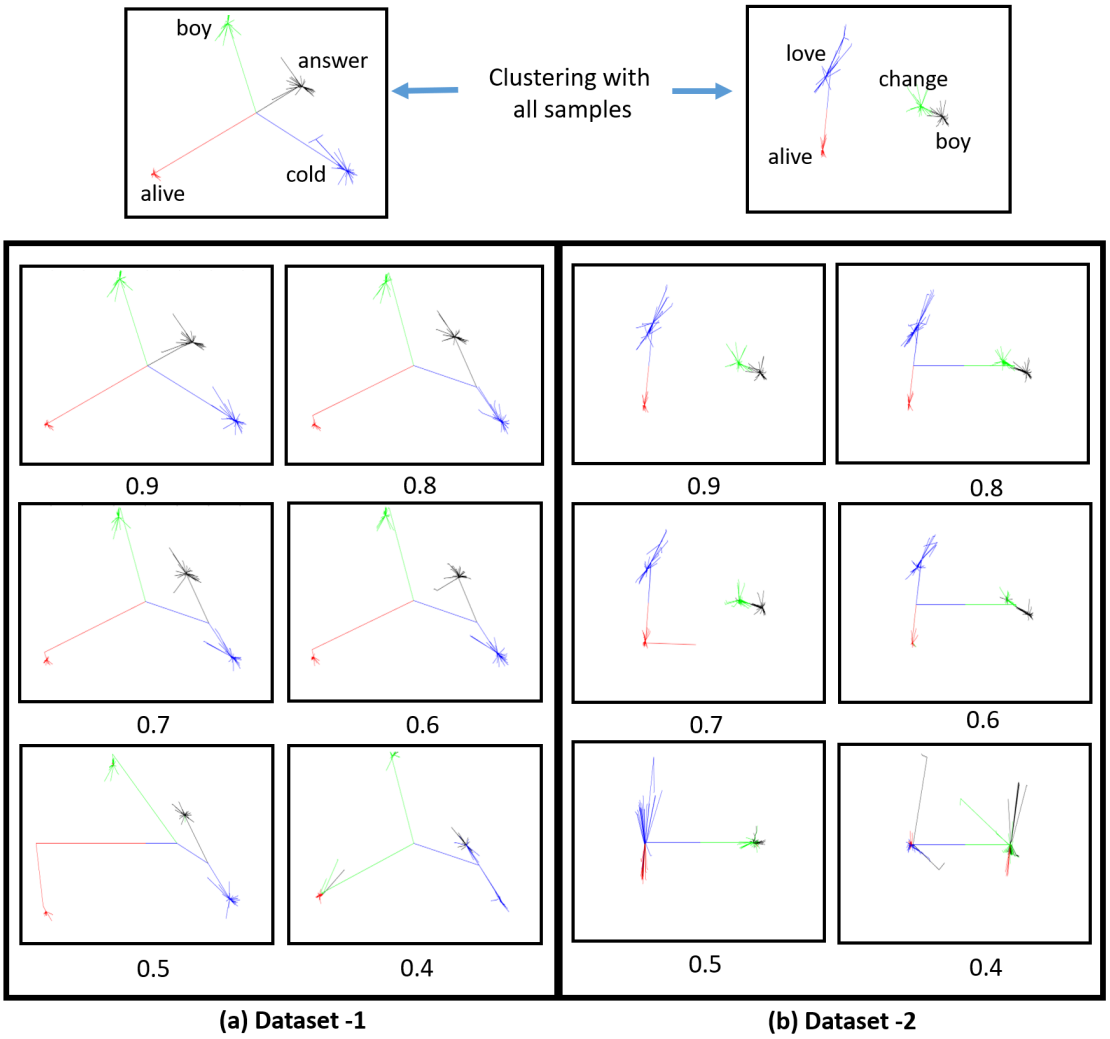


Figure 5.8: Clustering on subsets of words taken from the ASL dataset. 2 datasets have been shown here, with instances of 4 words in each case. Dataset-2 is more challenging to cluster in the presence of missing entries due to greater similarity between the 4 words, as indicated by a smaller separation distance. Dataset-1 is accurately clustered even for 40% missing entries, while Dataset-2 is accurately clustered for around 20% missing entries.

CHAPTER 6
SUMMARY & FUTURE DIRECTIONS

6.1 Summary

In this thesis, we have studied the problem of joint recovery of a group of signals from noisy or undersampled measurements, under different model assumptions. Traditionally the sparse and low-rank models had been considered for this problem. We extended the analysis to other models that have not been studied as extensively, yet are suitable for many real-world applications. We propose algorithms to solve the inverse problems, present theoretical guarantees for recovery under some model assumptions, and also demonstrate the algorithms on some practical problems. Our proposed algorithms make use of fusion penalties which enforce pairwise similarity between the different signals under consideration, thus exploiting the redundancies present in the dataset.

The first model that we consider is that of points lying on a low-dimensional manifold, embedded in high dimensional space. This model is satisfied by many datasets where each data point can be fully described by a low-dimensional parameter vector. Inspired by dimensionality reduction algorithms, we propose a signal recovery algorithm which enforces similarity between signals in local neighbourhoods of the manifold. We apply the proposed scheme to the problem of dynamic MRI reconstruction from few Fourier measurements. We propose a novel acquisition scheme which enables the detection of local neighbourhoods on the manifold. We get very promising results in free-breathing ungated cardiac and speech imaging applications, which indicate that the proposed scheme could serve as an alternative to clinical state-of-the art breath-held ECG-gated scans.

We also consider the problem of recovery of curves from few sampled points. For this purpose, we model the curves as the zero-level set of a trigonometric polynomial. We derive theoretical guarantees for the number of points required to uniquely re-

cover the curve, and note that it is dependent on the bandwidth of the underlying trigonometric polynomial. We apply the proposed technique to the reconstruction of DNA filaments from a few clicked points on noisy cryo-electron microscopy images. We extend the model to higher dimensions to enable the representation of surfaces. We present computationally efficient algorithms to recover points lying on this surface from their noisy or undersampled measurements. We demonstrate this algorithm on the recovery of simulated data, and also revisit the cardiac MRI reconstruction problem. We are able to recover better quality images in a shorter reconstruction time using this model.

Next, we consider the model of data arranged in clusters, with a few feature values unknown for each data point. Inspired by existing sum-of-norms clustering techniques, we propose an optimization problem to estimate the correct cluster centres even in the presence of missing entries. We present theoretical guarantees for its success, and note that the probability of success is greater for datasets with well-separated clusters, where the cluster memberships are not determined by only a few feature values. We present an efficient algorithm to solve a relaxation of this problem. We demonstrate the success of the proposed scheme on simulated data as well as real data such as the Wine and Australian Sign Language datasets.

Our proposed algorithms are general in nature, and we expect that they can be used in a variety of other applications, where the model assumptions are satisfied.

6.2 Future directions

Our algorithm for reconstruction of points lying on a manifold was inspired by existing dimensionality reduction algorithms and has been shown to work well on the problem of image reconstruction. However, we did not derive theoretical guarantees for the correct signal recovery. It would be interesting to study how the algorithm performance changes as a function of the properties of the manifold. Our acquisition scheme for dynamic MRI included special navigator signals which enabled the detec-

tion of local neighbourhoods on the manifold. While this worked well in practice, it would be interesting to find bounds for the accuracy of this estimate. Moreover, in applications other than MRI, different acquisition strategies would have to be explored in order to use the reconstruction algorithm.

We have considered the problem of recovery of curves from few sampled points. Since the proposed technique relies on the detection of the null-space vectors of a high-dimensional mapping of the sampled points, it may be highly sensitive to noise. Perhaps, an optimization algorithm penalizing the nuclear norm of the feature matrix would be more robust to noise for the curve recovery problem. Theoretical guarantees for this problem need to be studied in greater detail. Moreover, our current theoretical results using null-space methods have been derived only the 2D case, and their extension to higher dimensions is another direction that can be pursued. From our application to the recovery of DNA filaments, we have observed that a large number of points need to be clicked for good recovery. A potential direction to look at is to reduce this number by studying the effect of the location of the samples on the recovery guarantee.

We have presented an iterative algorithm for solving the clustering problem, which converges to a critical point of the original unconstrained optimization problem with saturating non-convex penalties. The connection between different initialization strategies and the critical point which is reached by the iterative algorithm could be studied in more detail. Moreover, the effect of moving from a constrained formulation to an unconstrained one, as well as the effect of the regularization parameter in the latter case could be studied theoretically.

We have shown promising results for the cardiac MRI reconstruction problem. The studies here were conducted in the 2D setting, i.e. slices were acquired one after the other. An alternative is to perform the imaging in the 3D setting. This enables the measurement of many clinically useful parameters. While the reconstruction

scheme can be extended to this setting fairly easily, more work needs to be done to design efficient acquisition schemes. Specifically, we are currently acquiring a few navigator signals every frame, which reduces the scan efficiency. While this did not cause problems in the 2D case, it might be difficult to achieve an acceptable temporal resolution in the 3D case. More efficient trajectories such as spirals might be better suited for this purpose.

APPENDIX A
PROOFS FOR CHAPTER 3

A.1 Proof of proposition 3.2.1

Proof. The polynomial $\psi(\mathbf{r})$ is represented in terms of its irreducible factors as:

$$\psi(\mathbf{r}) = \psi_1(\mathbf{r})\psi_2(\mathbf{r})\dots\psi_J(\mathbf{r}) \tag{A.1}$$

where the bandwidth of $\psi_j(\mathbf{r})$ is $K_1^j \times K_2^j$. Let $\eta(\mathbf{r})$ be another polynomial with bandwidth $K_1 \times K_2$ satisfying $\eta(\mathbf{x}_i) = 0$, for $i = 1, \dots, N$.

We consider the 2 polynomials $\psi_j(\mathbf{r})$ and $\eta(\mathbf{r})$. According to the required sampling condition, there are $N_j > (K_1 + K_2)(K_1^j + K_2^j)$ points satisfying $\psi_j(\mathbf{r}) = \eta(\mathbf{r}) = 0$. Thus, following Beźout's Theorem, $\psi_j(\mathbf{r})$ must be a factor of $\eta(\mathbf{r})$.

Following this line of reasoning for all factors $\{\psi^j\}$, it can be concluded that $\psi(\mathbf{r})$ is a factor of $\eta(\mathbf{r})$. However, since both $\psi(\mathbf{r})$ and $\eta(\mathbf{r})$ have the same bandwidth, the only possibility is that $\eta(\mathbf{r})$ is a scalar multiple of $\psi(\mathbf{r})$. Thus, the curve $\psi(\mathbf{r}) = 0$ can be uniquely recovered. The total number of points to be sampled is $N = \sum_{j=1}^{J} N_j > (K_1 + K_2)\sum_{j=1}^{J}(K_1^j + K_2^j)$. Using convolution properties, it can be concluded that $N > (K_1 + K_2)(K_1 + K_2 + 2(J-1))$. □

A.2 Proof of proposition 3.2.2

Proof. Following the steps of the proof for Proposition 1, we can conclude that $\psi(\mathbf{r})$ is a factor of $\psi'(\mathbf{r})$. Since $\Lambda \subset \Gamma$, it follows that $\psi'(\mathbf{r}) = \psi(\mathbf{r})\ \eta(\mathbf{r})$, where $\eta(\mathbf{r})$ is some arbitrary function such that $\psi'(\mathbf{r})$ is bandlimited to Γ. □

A.3 Proof of proposition 3.2.3

Proof. Let $\mathbf{c}$ be the minimal filter of band-width $|\Lambda|$, associated with the polynomial $\psi(\mathbf{r})$. We define the following filters supported in Γ for all $\mathbf{l} \in \Gamma : \Lambda$.

$$\mathbf{c_l}[\mathbf{k}] = \begin{cases} \mathbf{c}[\mathbf{k}-\mathbf{l}], & \text{if } \mathbf{k}-\mathbf{l} \in \Lambda. \\ 0, & \text{otherwise.} \end{cases} \tag{A.2}$$

$\mathbf{c_l}$ are the Fourier co-efficients of $\exp(j2\pi\mathbf{l}^T\mathbf{r})\psi(\mathbf{r})$, and are all null-space vectors of the feature matrix $\Phi_\Gamma(\mathbf{X})$. The number of such filters is $|\Gamma : \Lambda|$. Hence, we get the rank bound: $\text{rank}\,(\Phi_\Gamma(\mathbf{X})) \leq |\Gamma| - |\Gamma : \Lambda|$.

If the sampling conditions of Proposition 2 are satisfied, then all the polynomials corresponding to null-space vectors of Φ_Γ are of the form: $\psi'(\mathbf{r}) = \psi(\mathbf{r})\ \eta(\mathbf{r})$. Alternatively, in the Fourier domain, the filters are of the form:

$$\mathbf{c}'[\mathbf{k}] = \sum_{\mathbf{l} \in \Gamma:\Lambda} \mathbf{d_l} \mathbf{c_l}[\mathbf{k}] \tag{A.3}$$

where $\mathbf{d_l}$ are the Fourier co-efficients of the arbitrary polynomial $\eta(\mathbf{r})$. Thus, all the null-space filters can be represented in terms of the basis set $\{\mathbf{c_l}\}$. This leads to the relation: $\text{rank}\,(\Phi_\Gamma(\mathbf{X})) = |\Gamma| - |\Gamma : \Lambda|$. □

APPENDIX B
PROOFS FOR CHAPTER 5

B.1 Proof of lemma 5.2.1

Proof. Since $\mathbf{x}_1$ and $\mathbf{x}_2$ are in the same cluster, $\|\mathbf{x}_1 - \mathbf{x}_2\|_\infty \leq \epsilon$. For all the points in this particular cluster, let the p^{th} feature be bounded as: $f^p_{min} \leq \mathbf{x}(p) \leq f^p_{max}$. Then we can construct a vector $\mathbf{u}$, such that $\mathbf{u}(p) = \frac{1}{2}(f^p_{min} + f^p_{max})$. Now, since $f^p_{max} - f^p_{min} \leq \epsilon$, the following condition will be satisfied for this particular choice of $\mathbf{u}$:

$$\|\mathbf{x}_i - \mathbf{u}\|_\infty \;\leq\; \frac{\epsilon}{2}; \quad i = 1, 2 \tag{B.1}$$

From this, it follows trivially that the following will also hold:

$$\|\mathbf{S}_i\,(\mathbf{x}_i - \mathbf{u})\|_\infty \;\leq\; \frac{\epsilon}{2}; \quad i = 1, 2 \tag{B.2}$$

□

B.2 Lemma B.2.1

Lemma B.2.1. *Consider any pair of points $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^P$ observed by sampling matrices $\mathbf{S}_1 = \mathbf{S}_{\mathcal{I}_1}$ and $\mathbf{S}_2 = \mathbf{S}_{\mathcal{I}_2}$, respectively. We assume the set of common indices ($\omega \coloneqq \mathcal{I}_1 \cap \mathcal{I}_2$) to be of size $q = |\mathcal{I}_1 \cap \mathcal{I}_2|$. Then, for some $0 < t < \frac{q}{P}$, the following result holds true regarding the partial distance $\|\mathbf{y}_\omega\|_2 = \|\mathbf{S}_{\mathcal{I}_1 \cap \mathcal{I}_2}\,(\mathbf{x}_1 - \mathbf{x}_2)\,\|_2$:*

$$\mathbb{P}\left(\|\mathbf{y}_\omega\|_2^2 \leq \left(\frac{q}{P} - t\right)\|\mathbf{y}\|_2^2\right) \leq e^{-\frac{2t^2P^2}{q\mu_0^2}} \tag{B.3}$$

Proof. We use some ideas for bounding partial distances from Lemma 3 of [24]. We rewrite the partial distance $\|\mathbf{y}_\omega\|_2^2$ as the sum of q variables drawn uniformly at random from $\{y_1^2, y_2^2, \ldots, y_P^2\}$. By replacing a particular variable in the summation by another one, the value of the sum changes by at most $\|\mathbf{y}\|_\infty^2$. Applying McDiarmid's Inequality,

we get:

$$\mathbb{P}\left(E(\|\mathbf{y}_\omega\|_2^2) - \|\mathbf{y}_\omega\|_2^2 \geq c\right) \leq e^{-\frac{2c^2}{\sum_{i=1}^{q}\|\mathbf{y}\|_\infty^4}} = e^{-\frac{2c^2}{q\|\mathbf{y}\|_\infty^4}} \tag{B.4}$$

From our assumptions, we have $E(\|\mathbf{y}_\omega\|_2^2) = \frac{q}{P}\|\mathbf{y}\|_2^2$. We also have $\frac{\|\mathbf{y}\|_2^2}{\|\mathbf{y}\|_\infty^2} \geq \frac{P}{\mu_0}$ by (5.6). We now substitute $c = t\|\mathbf{y}\|_2^2$, where $0 < t < \frac{q}{P}$. Using the results above, we simplify expression (B.4) as:

$$\begin{aligned}\mathbb{P}\left(\|\mathbf{y}_\omega\|_2^2 \leq \left(\frac{q}{P} - t\right)\|\mathbf{y}\|_2^2\right) &\leq e^{-\frac{2t^2\|\mathbf{y}\|_2^4}{q\|\mathbf{y}\|_\infty^4}} \\ &\leq e^{-\frac{2t^2P^2}{q\mu_0^2}}\end{aligned} \tag{B.5}$$

□

B.3 Proof of lemma 5.2.2

Proof. We will use proof by contradiction. Specifically, we consider two points $\mathbf{x}_1$ and $\mathbf{x}_2$ belonging to different clusters and assume that there exists a point $\mathbf{u}$ that satisfies:

$$\|\mathbf{S}_i\left(\mathbf{x}_i - \mathbf{u}\right)\|_\infty \quad \leq \quad \frac{\epsilon}{2}; i = 1, 2 \tag{B.6}$$

We now show that the above assumption is violated with high probability. Following the notation of Lemma B.2.1, we denote the difference between the vectors by $\mathbf{y} = \mathbf{x}_1 - \mathbf{x}_2$ and the partial distances by:

$$\|\mathbf{y}_\omega\|_2 = \|\mathbf{S}_{\mathcal{I}_1 \cap \mathcal{I}_2}\ \left(\mathbf{x}_1 - \mathbf{x}_2\right)\|_2 \tag{B.7}$$

Using (B.6) and applying triangle inequality, we obtain $\|\mathbf{y}_\omega\|_\infty \leq \epsilon$, which translates to $\|\mathbf{y}_\omega\|_2 \leq \epsilon\sqrt{q}$, where $q = |\mathcal{I}_1 \cap \mathcal{I}_2|$ is the number of commonly observed locations. We need to show that with high probability, the partial distances satisfy:

$$\|\mathbf{y}_\omega\|_2^2 > \epsilon^2 q \tag{B.8}$$

which will contradict (B.6). We first focus on finding a lower bound for q. Using the Chernoff bound and setting $\mathbb{E}(q) = p_0^2 P$, we have:

$$\mathbb{P}\left(q \geq \frac{p_0^2 P}{2}\right) > 1 - \gamma_0 \tag{B.9}$$

where $\gamma_0 = (\frac{e}{2})^{-\frac{p_0^2 P}{2}}$. Thus, we can assume that $q \geq \frac{p_0^2 P}{2}$ with high probability.

Using Lemma B.2.1, we have the following result for the partial distances:

$$\mathbb{P}\left(\|\mathbf{y}_\omega\|_2^2 \leq \left(\frac{q}{P} - t\right)\|\mathbf{y}\|_2^2\right) \leq e^{-\frac{2t^2P^2}{q\mu_0^2}} \tag{B.10}$$

Since $\mathbf{x}_1$ and $\mathbf{x}_2$ are in different clusters, we have $\|\mathbf{y}\|_2 \geq \delta$. We will now determine the value of t for which the above upper bound will equal the RHS of (B.8):

$$\left(\frac{q}{P} - t\right)\|\mathbf{y}\|_2^2 = \epsilon^2 q \tag{B.11}$$

or equivalently:

$$t = \frac{q}{P} - \frac{\epsilon^2 q}{\|\mathbf{y}\|_2^2} \geq \frac{q}{P} - \frac{\epsilon^2 q}{\delta^2} = \frac{q}{P}(1 - \kappa^2) \tag{B.12}$$

Since $t > 0$, we require $\kappa < 1$, where $\kappa = \frac{\epsilon\sqrt{P}}{\delta}$. Using the above, we get the following bound if we assume that $q \geq \frac{p_0^2 P}{2}$:

$$\frac{t^2}{q} \geq \frac{q}{P^2}(1 - \kappa^2)^2 \geq \frac{p_0^2}{2P}(1 - \kappa^2)^2 \tag{B.13}$$

We now obtain the following probability bound for any $q \geq \frac{p_0^2 P}{2}$:

$$\begin{aligned}
\mathbb{P}\left(\|\mathbf{y}_\omega\|^2 > \epsilon^2 q\right) &\geq 1 - e^{-\frac{2t^2P^2}{q\mu_0^2}} \\
&\geq 1 - e^{-\frac{p_0^2 P(1-\kappa^2)^2}{\mu_0^2}} \\
&= 1 - \delta_0
\end{aligned} \tag{B.14}$$

Combining (B.9) and (B.14), the probability for (B.6) to hold is $\leq 1-(1-\gamma_0)(1-\delta_0) = \beta_0$.

□

B.4 Proof of lemma 5.2.3

Proof. We construct a graph where each point $\mathbf{x}_i$ is represented by a node. Lemma 5.2.1 implies that a pair of points belonging to the same cluster can yield the same $\mathbf{u}$ in a feasible solution with probability 1. Hence, we will assume that there exists an edge between two nodes from the same cluster with probability 1. Lemma 5.2.2 indicates that a pair of points belonging to different clusters can yield the same $\mathbf{u}$ in a feasible solution with a low probability of β_0. We will assume that there exists an edge between two nodes from different clusters with probability β_0. We will now evaluate the probability that there exists a fully-connected sub-graph of size M, where all the nodes have not been taken from the same cluster. We will follow a methodology similar to [53], which gives an expression for the probability distribution of the maximal clique (i.e. largest fully connected sub-graph) size in a random graph. Unlike the proof in [53], in our graph every edge is not present with equal probability.

We define the following random variables:

- $t \coloneqq$ Size of the largest fully connected sub-graph containing nodes from more than 1 cluster
- $n \coloneqq$ Number of M membered complete sub-graphs containing nodes from more than 1 cluster

Our graph can have an M membered clique iff n is non-zero. Thus, we have:

$$\mathbb{P}(t \geq M) = \mathbb{P}(n \neq 0) \tag{B.15}$$

Since the distribution of n is restricted only to the non-negative integers, it can be

seen that:

$$\mathbb{P}\left(n \neq 0\right) \leq E(n) \tag{B.16}$$

Combining the above 2 results, we get:

$$\mathbb{P}\left(t \geq M\right) \leq E(n) \tag{B.17}$$

Let us consider the formation of a particular clique of size M using $m_1, m_2, \ldots, m_K$ nodes from clusters $C_1, C_2, \ldots, C_K$ respectively such that $\sum_{j=1}^{K} m_j = M$, and at least 2 of the variables $\{m_j\}$ are non-zero. The number of ways to choose such a collection of nodes is: $\prod_j \binom{M}{m_j}$. In order to form a solution $\{m_j\}$, we need $\frac{1}{2}(M^2 - \sum_j m_j^2)$ inter-cluster edges to be present. We recall that each of these edges is present with probability β_0. Thus, the probability that such a collection of nodes forms a clique is $\beta_0^{\frac{1}{2}(M^2 - \sum_j m_j^2)}$. This gives the following result:

$$E(N) = \sum_{\{m_j\} \in \mathcal{S}} \beta_0^{\frac{1}{2}(M^2 - \sum_j m_j^2)} \prod_j \binom{M}{m_j} = \eta_0 \tag{B.18}$$

where $\mathcal{S}$ is the set of all sets of positive integers $\{m_j\}$ such that: $2 \leq \mathcal{U}(\{m_j\}) \leq K$ and $\sum_j m_j = M$. Here, the function $\mathcal{U}$ counts the number of non-zero elements in a set. Thus, we have:

$$\mathbb{P}\left(t \geq M\right) \leq \eta_0 \tag{B.19}$$

This proves that with probability $\geq 1 - \eta_0$, a set of points of cardinality $\geq M$ not all belonging to the same cluster cannot all have equal cluster-centre estimates.

□

B.5 Proof of theorem 5.2.4

Proof. Lemma 5.2.1 indicates that fully connected original clusters with size M are likely with probability 1, while Lemma 5.2.3 shows that the size of misclassified large

clusters cannot exceed $M-1$ with very high probability. These results enable us to re-express the optimization problem (5.8) as a simpler maximization problem. We will then show that with high probability, any feasible solution other than the ground-truth solution results in a cost higher than the ground-truth solution.

Let a candidate solution have k groups of sizes $M_1, M_2, \ldots, M_k$ respectively. The centre estimates for all points within a group are equal. These are different from the centre estimates of other groups. Without loss of generality, we will assume that at most K of these groups each have points belonging to only a single ground-truth cluster, i.e. they are "pure". The rest of the clusters in the candidate solution are "mixed" clusters. If we have a candidate solution with greater than K pure clusters, then they can always be merged to form K pure clusters; the merged solution will always result in a lower cost.

The objective function in (5.8) can thus be rewritten as:

$$\begin{aligned}\sum_{i=1}^{KM}\sum_{j=1}^{KM}\|\mathbf{u}_i-\mathbf{u}_j\|_{2,0} &= \sum_{i=1}^{k} M_i(KM-M_i)\\ &= K^2M^2-\sum_{i=1}^{k}M_i^2\end{aligned} \tag{B.20}$$

Since we assume that the first K clusters are pure, therefore they have a size $0 \leq M_i \leq M$, $i=1,\ldots,K$. The remaining clusters are mixed and have size $\leq M-1$ with probability $\geq 1-\eta_0$. Hence, we have the constraints $0 \leq M_i \leq (M-1)$, $i=K+1,\ldots,k$. We also have a constraint on the total number of points, i.e. $\sum_{i=1}^{k} M_i = KM$. Thus, the problem (5.8) can be rewritten as the constrained optimization

problem:

$$\begin{aligned} \{M_i^*, k^*\} = \max_{\{M_i\},k} & \sum_{i=1}^{k} M_i^2 \\ \text{s.t. } & 0 \leq M_i \leq M, i = 1, \ldots, K \\ & 0 \leq M_i \leq M-1, i = K+1, \ldots, k \\ & \sum_{i=1}^{k} M_i = KM \end{aligned} \tag{B.21}$$

Note that we cannot have $k < K$, with probability $\geq 1 - \eta_0$, since that involves a solution with cluster size $> M$. We can evaluate the best solution $\{M_i^*\}$ for each possible value of k in the range $K \leq k \leq MK$. Then we can compare these solutions to get the solution with the highest cost. We note that the feasible region is a polyhedron and the objective function is convex. Thus, for each value of k, we only need to check the cost at the vertices of the polyhedron formed by the constraints, since the cost at all other points in the feasible region will be lower. The vertex points are formed by picking $k-1$ out of the k box constraints and setting M_i to be equal to one of the 2 possible extremal values. We note that all the vertex points have either K or $K+1$ non-zero values. As a simple example, if we choose $M = 10$ and $K = 4$, then the vertex points of the polyhedron (corresponding to different solutions $\{M_i\}$) are given by all possible permutations of the following:

- $(10, 10, 10, 10, 0, 0 \ldots 0)$: 4 clusters
- $(10, 10, 10, 0, 1, 9, 0 \ldots 0)$: 5 clusters
- $(10, 10, 0, 0, 2, 9, 9, 0 \ldots 0)$: 5 clusters
- $(10, 0, 0, 0, 3, 9, 9, 9, 0 \ldots 0)$: 5 clusters
- $(0, 0, 0, 0, 4, 9, 9, 9, 9, 0 \ldots 0)$: 5 clusters

In the general case the vertices are given by permutations of the following:

- $(M, M, \ldots, M, 0, 0 \ldots 0)$: K clusters
- $(M, M, \ldots, 0, 0, 1, M-1, 0 \ldots 0)$: $K+1$ clusters
- $(M, M, \ldots, 0, 0, 2, M-1, M-1 \ldots 0)$: $K+1$ clusters
- ...
- $(0, 0, \ldots 0, K, M-1, M-1 \ldots M-1, 0)$: $K+1$ clusters

Now, it is easily checked that the 1^{st} candidate solution in the list (which is also the ground-truth solution) has the maximum cost. Mixed clusters with size $> M-1$ cannot be formed with probability $> 1-\eta_0$. Thus, with the same probability, the solution to the optimization problem (5.8) is identical to the ground-truth clustering. This concludes the proof of the theorem.

□

B.6 Upper bound for η_0 in the 2-cluster case

Proof. We introduce the following notation:

1. $F(i) = i(M-i)\log\beta_0$, for $i \in [1, M-1]$.
2. $G(i) = 2[\log\Gamma(M+1) - \log\Gamma(i+1) - \log\Gamma(M-i+1)]$, for $i \in [1, M-1]$ where Γ is the Gamma function.

We note that both the functions F and G are symmetric about $i = \frac{M}{2}$, and have unique minimum and maximum respectively for $i = \frac{M}{2}$. We will show that the maximum for the function $F+G$ is achieved at the points $i = 1, M-1$. We note that:

$$G'(i) = -2[\Psi(i+1) - \Psi(M-i+1)] \tag{B.22}$$

where Ψ is the digamma function, defined as the log derivative of the Γ function. We now use the expansion:

$$\Psi(i+1) = \log i + \frac{1}{2i} \tag{B.23}$$

Substituting, we get:

$$G'(i) = -2\left[\log\frac{i}{M-i} + \frac{M-2i}{2i(M-i)}\right] \tag{B.24}$$

We also have:

$$F'(i) = (M-2i)\log\beta_0 \tag{B.25}$$

Adding, we get:

$$\begin{aligned} F'(i) + G'(i) = (M-2i)(\log\beta_0 - \frac{1}{i(M-i)}) \\ - 2\log\frac{i}{(M-i)}) \end{aligned} \tag{B.26}$$

Now, in order to ensure that $F'(i) + G'(i) \leq 0$, we have to arrive at conditions such that:

$$\log\beta_0 \leq \frac{1}{i(M-i)} + \frac{2}{M-2i}\log\frac{i}{M-i} \tag{B.27}$$

Since the RHS is monotonically increasing in the interval $i \in [1, \frac{M}{2} - 1]$ the above condition reduces to:

$$\log\beta_0 \leq \frac{1}{M-1} + \frac{2}{M-2}\log\frac{1}{M-1} \tag{B.28}$$

Under the above condition, for all $i \in [1, \frac{M}{2}]$:

$$F'(i) + G'(i) \leq 0 \tag{B.29}$$

Thus, the function $F + G$ reaches its maxima at the extremal points given by $i = 1, M-1$. For positive integer values of i, i.e. $i \in \{1, 2, \ldots, M-1\}$:

$$F(i) + G(i) = \log[\beta_0^{i(M-i)} \binom{M}{i}^2] \tag{B.30}$$

Thus, the function $\beta_0^{i(M-i)}\binom{M}{i}^2$ also reaches its maxima at $i = 1, M-1$. This maximum value is given by: $\beta_0^{M-1}M^2$. This gives the following upper bound for η_0:

$$\begin{aligned} \eta_0 &\leq \sum_{i=1}^{M-1} [\beta_0^{M-1} M^2] \\ &= M^2(M-1)\beta_0^{M-1} \\ &\leq M^3 \beta_0^{M-1} \\ &= \eta_{0,\text{approx}} \end{aligned} \tag{B.31}$$

□

B.7 Proof of theorem 5.2.5

Proof. We consider any two points $\mathbf{x}_1$ and $\mathbf{x}_2$ that are in different clusters. Let us assume that there exists some $\mathbf{u}$ satisfying the data consistency constraint:

$$\|\mathbf{x}_i - \mathbf{u}\|_\infty \leq \epsilon/2, \quad i = 1, 2. \tag{B.32}$$

Using the triangle inequality, we have $\|\mathbf{x}_1 - \mathbf{x}_2\|_\infty \leq \epsilon$ and consequently, $\|\mathbf{x}_1 - \mathbf{x}_2\|_2 \leq \epsilon\sqrt{P}$. However, if we have a large inter-cluster separation $\delta > \epsilon\sqrt{P}$, then this is not possible.

Thus, if $\delta > \epsilon\sqrt{P}$, then points in different clusters cannot be misclassified to a single cluster. Among all feasible solutions, clearly the solution to problem (5.25) with the minimum cost is the one where all points in the same cluster merge to the same $\mathbf{u}$. Thus, $\kappa < 1$ ensures that we will have the correct clustering. □

B.8 Proof of lemma 5.2.6

Proof. The idea is similar to that in Theorem 5.2.5. We will show that with high probability two points $\mathbf{x}_1$ and $\mathbf{x}_2$ that are in different clusters satisfy $\|\mathbf{x}_1 - \mathbf{x}_2\|_2 > \epsilon\sqrt{P}$ with high probability, which implies that (5.29) is violated.

Let points in C_1 and C_2 follow uniform random distributions in $\mathbb{R}^P$ with centres $\mathbf{c}_1$ and $\mathbf{c}_2$ respectively. The expected distance between $\mathbf{x}_1 \in \mathcal{C}_1$ and $\mathbf{x}_2 \in \mathcal{C}_2$ is given by:

$$\begin{aligned} E(\|\mathbf{x}_1 - \mathbf{x}_2\|_2^2) &= \frac{1}{\epsilon^2} \sum_{p=1}^{P} \int_{\mathbf{c}_1^p - \frac{\epsilon}{2}}^{\mathbf{c}_1^p + \frac{\epsilon}{2}} \int_{\mathbf{c}_2^p - \frac{\epsilon}{2}}^{\mathbf{c}_2^p + \frac{\epsilon}{2}} (\mathbf{x}_1^p - \mathbf{x}_2^p)^2 d\mathbf{x}_1^p d\mathbf{x}_2^p \\ &= \|\mathbf{c}_1 - \mathbf{c}_2\|_2^2 + \frac{P}{6}\epsilon^2 \\ &= c_{12}^2 + \frac{P}{6}\epsilon^2 \end{aligned} \tag{B.33}$$

where $\mathbf{c}_i^p$ and $\mathbf{x}_i^p$ are the p^{th} features of $\mathbf{c}_i$ and $\mathbf{x}_i$ respectively, and $c_12 = \|\mathbf{c}_1 - \mathbf{c}_2\|_2$. Let $c_i = |\mathbf{c}_1^i - \mathbf{c}_2^i|$, for $i = 1, 2, \ldots, P$. Using Mcdiarmid's inequality:

$$\begin{aligned} &\mathbb{P}\left(\|\mathbf{x}_1 - \mathbf{x}_2\|_2^2 \leq E(\|\mathbf{x}_1 - \mathbf{x}_2\|_2^2) - t\right) \\ &\leq e^{-\frac{2t^2}{\sum_{i=1}^{P} |(c_i+\epsilon)^2 - (c_i-\epsilon)^2|^2}} \\ &= e^{-\frac{t^2}{8\epsilon^2 c_{12}^2}} \end{aligned} \tag{B.34}$$

Let $t = E(\|\mathbf{x}_1 - \mathbf{x}_2\|_2^2) - P\epsilon^2$. Then we have:

$$\mathbb{P}\left(\|\mathbf{x}_1 - \mathbf{x}_2\|_2 \leq \epsilon\sqrt{P}\right) \leq e^{-\frac{(c_{12}^2 - \frac{5P}{6}\epsilon^2)^2}{8\epsilon^2 c_{12}^2}} \tag{B.35}$$

We note that the RHS above is a decreasing function of c_{12}. Thus, we consider some $c \leq c_{12}$, such that c is the minimum distance between any 2 cluster centres in the

dataset. We then have the following bound:

$$\mathbb{P}\left(\|\mathbf{x}_1 - \mathbf{x}_2\|_2 \leq \epsilon\sqrt{P}\right) \leq e^{-\frac{(c^2 - \frac{5P}{6}\epsilon^2)^2}{8\epsilon^2 c^2}} \tag{B.36}$$

To ensure $t > 0$, we require: $c > \sqrt{\frac{5P}{6}}\epsilon$, or equivalently, $\kappa' = \frac{\epsilon\sqrt{P}}{c} < \sqrt{\frac{6}{5}}$.

We now get the probability bound:

$$\mathbb{P}\left(\|\mathbf{x}_1 - \mathbf{x}_2\|_2 \leq \epsilon\sqrt{P}\right) \leq e^{-\frac{P(1-\frac{5}{6}\kappa'^2)^2}{8\kappa'^2}} = \beta_1 \tag{B.37}$$

Thus, (5.29) is violated with probability exceeding $1 - \beta_1$. □

REFERENCES

[1] P. Awasthi, A. S. Bandeira, M. Charikar, R. Krishnaswamy, S. Villar, and R. Ward, "Relax, no need to round: Integrality of clustering formulations," in *Proceedings of the 2015 Conference on Innovations in Theoretical Computer Science*. ACM, 2015, pp. 191–200.

[2] A. Balachandrasekaran and M. Jacob, "Novel structured low-rank algorithm to recover spatially smooth exponential image time series," in *Biomedical Imaging (ISBI 2017), 2017 IEEE 14th International Symposium on*. IEEE, 2017, pp. 1–4.

[3] A. Balachandrasekaran, V. Magnotta, and M. Jacob, "Recovery of damped exponentials using structured low rank matrix completion," *IEEE transactions on medical imaging*, vol. 36, no. 10, pp. 2087–2098, 2017.

[4] M. Belkin, "Problems of learning on manifolds," Ph.D. dissertation, 2003.

[5] M. Belkin and P. Niyogi, "Laplacian eigenmaps and spectral techniques for embedding and clustering," in *Advances in neural information processing systems*, 2002, pp. 585–591.

[6] R. M. Bell, Y. Koren, and C. Volinsky, "The bellkor 2008 solution to the netflix prize," *Statistics Research Department at AT&T Research*, 2008.

[7] K. K. Bhatia, J. Caballero, A. N. Price, J. V. Hajnal, and D. Rueckert, "Fast reconstruction of highly-undersampled dynamic MRI using random sampling and manifold interpolation," in *International Society on Magnetic Resonance in Medicine 2015*.

[8] P. S. Bradley, O. L. Mangasarian, and W. N. Street, "Clustering via concave minimization," in *Advances in neural information processing systems*, 1997, pp. 368–374.

[9] J. M. Brick and G. Kalton, "Handling missing data in survey research," *Statistical methods in medical research*, vol. 5, no. 3, pp. 215–238, 1996.

[10] E. J. Candès and C. Fernandez-Granda, "Towards a mathematical theory of super-resolution," *Communications on Pure and Applied Mathematics*, vol. 67, no. 6, pp. 906–956, 2014.

[11] E. J. Candès and B. Recht, "Exact matrix completion via convex optimization," *Foundations of Computational mathematics*, vol. 9, no. 6, p. 717, 2009.

[12] R. Chartrand and W. Yin, "Iteratively reweighted algorithms for compressive sensing," in *Acoustics, speech and signal processing, 2008. ICASSP 2008. IEEE international conference on.* IEEE, 2008, pp. 3869–3872.

[13] G. K. Chen, E. C. Chi, J. M. O. Ranola, and K. Lange, "Convex clustering: An attractive alternative to hierarchical clustering," *PLoS Comput Biol*, vol. 11, no. 5, p. e1004228, 2015.

[14] X. Chen, M. Usman, C. F. Baumgartner, D. R. Balfour, P. K. Marsden, A. J. Reader, C. Prieto, and A. P. King, "High-resolution self-gated dynamic abdominal mri using manifold alignment," *IEEE transactions on medical imaging*, vol. 36, no. 4, pp. 960–971, 2017.

[15] E. C. Chi and K. Lange, "Splitting methods for convex clustering," *Journal of Computational and Graphical Statistics*, vol. 24, no. 4, pp. 994–1013, 2015.

[16] J. T. Chi, E. C. Chi, and R. G. Baraniuk, "k-pod: A method for k-means clustering of missing data," *The American Statistician*, vol. 70, no. 1, pp. 91–99, 2016.

[17] A. G. Christodoulou, C. Brinegar, J. P. Haldar, H. Zhang, Y.-J. L. Wu, L. M. Foley, T. K. Hitchens, Q. Ye, C. Ho, and Z.-P. Liang, "High-resolution cardiac mri using partially separable functions and weighted spatial smoothness regularization," in *Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC).* IEEE, 2010, pp. 871–874.

[18] I. Daubechies, R. DeVore, M. Fornasier, and C. S. Güntürk, "Iteratively reweighted least squares minimization for sparse recovery," *Communications on pure and applied mathematics*, vol. 63, no. 1, pp. 1–38, 2010.

[19] M. C. De Souto, P. A. Jaskowiak, and I. G. Costa, "Impact of missing data imputation methods on gene expression clustering and classification," *BMC bioinformatics*, vol. 16, no. 1, p. 64, 2015.

[20] J. K. Dixon, "Pattern recognition with partly missing data," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 9, no. 10, pp. 617–621, 1979.

[21] D. L. Donoho, "Compressed sensing," *IEEE Transactions on information theory*, vol. 52, no. 4, pp. 1289–1306, 2006.

[22] A. Eftekhari and M. B. Wakin, "New analysis of manifold embeddings and signal recovery from compressive measurements," *Applied and Computational Harmonic Analysis*, vol. 39, no. 1, pp. 67–109, 2015.

[23] E. Elhamifar, "High-rank matrix completion and clustering under self-expressive models," in *Advances in Neural Information Processing Systems*, 2016, pp. 73–81.

[24] B. Eriksson, L. Balzano, and R. D. Nowak, "High-rank matrix completion and subspace clustering with missing data," *CoRR*, vol. abs/1112.5629, 2011. [Online]. Available: http://arxiv.org/abs/1112.5629

[25] L. Feng, L. Axel, H. Chandarana, K. T. Block, D. K. Sodickson, and R. Otazo, "Xd-grasp: Golden-angle radial mri with reconstruction of extra motion-state dimensions using compressed sensing," *Magnetic resonance in medicine*, vol. 75, no. 2, pp. 775–788, 2016.

[26] L. Feng, D. Sodickson, and R. Otazo, "A Robust and Automatic Cardiac and Respiratory Motion Detection Framework for Self-Navigated Radial MRI," in *International Society on Magnetic Resonance in Medicine 2014*.

[27] J. Friedman, T. Hastie, H. Höfling, R. Tibshirani, *et al.*, "Pathwise coordinate optimization," *The Annals of Applied Statistics*, vol. 1, no. 2, pp. 302–332, 2007.

[28] S. B. Gay, C. L. Sistrom, C. A. Holder, and P. M. Suratt, "Breath-Holding Capability of Adults: Implications for Spiral Computed Tomography, Fast-Acquisition Magnetic Resonance Imaging, and Angiography," *Investigative Radiology*, vol. 29, 1994.

[29] G. Gilboa and S. Osher, "Nonlocal operators with applications to image processing," *Multiscale Modeling & Simulation*, vol. 7, no. 3, pp. 1005–1028, 2008.

[30] J. P. Haldar, "Low-rank modeling of local k-space neighborhoods (loraks) for constrained mri," *IEEE transactions on medical imaging*, vol. 33, no. 3, pp. 668–681, 2014.

[31] J. P. Haldar and J. Zhuo, "P-loraks: Low-rank modeling of local k-space neighborhoods with parallel imaging data," *Magnetic resonance in medicine*, vol. 75, no. 4, pp. 1499–1514, 2016.

[32] J. A. Hartigan and M. A. Wong, "Algorithm as 136: A k-means clustering algorithm," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, vol. 28, no. 1, pp. 100–108, 1979.

[33] R. J. Hathaway and J. C. Bezdek, "Fuzzy c-means clustering of incomplete data," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 31, no. 5, pp. 735–744, 2001.

[34] T. D. Hocking, A. Joulin, F. Bach, and J.-P. Vert, "Clusterpath an algorithm for clustering using convex fusion penalties," in *28th international conference on machine learning*, 2011, p. 1.

[35] H. Hoefling, "A path algorithm for the fused lasso signal approximator," *Journal of Computational and Graphical Statistics*, vol. 19, no. 4, pp. 984–1006, 2010.

[36] L. Hunt and M. Jorgensen, "Mixture model clustering for mixed data with missing information," *Computational Statistics & Data Analysis*, vol. 41, no. 3, pp. 429–440, 2003.

[37] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: a review," *ACM computing surveys (CSUR)*, vol. 31, no. 3, pp. 264–323, 1999.

[38] K. H. Jin, D. Lee, and J. C. Ye, "A general framework for compressed sensing and parallel mri using annihilating filter based low-rank hankel matrix," *IEEE Transactions on Computational Imaging*, vol. 2, no. 4, pp. 480–495, 2016.

[39] H. Jung, K. Sung, K. S. Nayak, E. Y. Kim, and J. C. Ye, "k-t FOCUSS: A general compressed sensing framework for high resolution dynamic MRI," *Magnetic Resonance in Medicine*, vol. 61, no. 1, pp. 103–116, 2009.

[40] M. W. Kadous *et al.*, *Temporal classification: Extending the classification paradigm to multivariate time series*. University of New South Wales, 2002.

[41] V. Kalofolias, "How to learn a graph from smooth signals," in *Proc. Int. Conf. Artificial Intelligence and Statistics*, vol. 51, 2016, pp. 920–929.

[42] F. Knoll, K. Bredies, T. Pock, and R. Stollberger, "Second order total generalized variation (TGV) for MRI," *Magnetic resonance in medicine*, vol. 65, no. 2, pp. 480–491, 2011.

[43] S. R. Land and J. H. Friedman, "Variable fusion: A new adaptive signal regression method," Technical Report 656, Department of Statistics, Carnegie Mellon University Pittsburgh, Tech. Rep., 1997.

[44] D. Lee, K. H. Jin, E. Y. Kim, S.-H. Park, and J. C. Ye, "Acceleration of mr parameter mapping using annihilating filter-based low rank hankel matrix (aloha)," *Magnetic resonance in medicine*, vol. 76, no. 6, pp. 1848–1864, 2016.

[45] Z.-P. Liang, "Spatiotemporal imagingwith partially separable functions," in *Biomedical Imaging: From Nano to Macro, 2007. ISBI 2007. 4th IEEE International Symposium on*. IEEE, 2007, pp. 988–991.

[46] M. Lichman, "UCI machine learning repository," 2013. [Online]. Available: http://archive.ics.uci.edu/ml

[47] T. I. Lin, J. C. Lee, and H. J. Ho, "On fast supervised learning for normal mixture models with missing information," *Pattern Recognition*, vol. 39, no. 6, pp. 1177–1187, 2006.

[48] S. G. Lingala, Y. Hu, E. DiBella, and M. Jacob, "Accelerated dynamic mri exploiting sparsity and low-rank structure: kt slr," *Medical Imaging, IEEE Transactions on*, vol. 30, no. 5, pp. 1042–1054, 2011.

[49] S. G. Lingala and M. Jacob, "Blind compressive sensing dynamic mri," *IEEE transactions on medical imaging*, vol. 32, no. 6, pp. 1132–1145, 2013.

[50] J. Liu, L. Yuan, and J. Ye, "An efficient algorithm for a class of fused lasso problems," in *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2010, pp. 323–332.

[51] M. Lustig, J. Santos, D. Donoho, and J. Pauly, "k-t SPARSE: High frame rate dynamic MRI exploiting spatio-temporal sparsity," in *International Society on Magnetic Resonance in Medicine 2006*.

[52] J. MacQueen *et al.*, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, vol. 1, no. 14. Oakland, CA, USA., 1967, pp. 281–297.

[53] D. W. Matula, *The largest clique size in a random graph*. Department of Computer Science, Southern Methodist University, 1976.

[54] Y. Q. Mohsin, S. G. Lingala, E. DiBella, and M. Jacob, "Accelerated dynamic mri using patch regularization for implicit motion compensation," *Magnetic Resonance in Medicine*, vol. 77, no. 3, pp. 1238–1248, 2017. [Online]. Available: http://dx.doi.org/10.1002/mrm.26215

[55] Y. Q. Mohsin, G. Ongie, and M. Jacob, "Iterative shrinkage algorithm for patch-smoothness regularized medical image recovery," *IEEE transactions on medical imaging*, vol. 34, no. 12, pp. 2417–2428, 2015.

[56] U. Nakarmi, Y. Wang, J. Lyu, and L. Ying, "Dynamic magnetic resonance imaging using compressed sensing with self-learned nonlinear dictionary (NL-D)," in *2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI)*, April 2015, pp. 331–334.

[57] S. Narayanan, A. Toutios, V. Ramanarayanan, A. Lammert, J. Kim, S. Lee, K. Nayak, Y.-C. Kim, Y. Zhu, L. Goldstein, *et al.*, "Real-time magnetic resonance imaging and electromagnetic articulography database for speech production research (TC)," *The Journal of the Acoustical Society of America*, vol. 136, no. 3, pp. 1307–1311, 2014.

[58] A. Y. Ng, M. I. Jordan, Y. Weiss, *et al.*, "On spectral clustering: Analysis and an algorithm," in *NIPS*, vol. 14, no. 2, 2001, pp. 849–856.

[59] P. Niyogi, S. Smale, and S. Weinberger, "Finding the homology of submanifolds with high confidence from random samples," *Discrete & Computational Geometry*, vol. 39, no. 1-3, pp. 419–441, 2008.

[60] G. Ongie, S. Biswas, and M. Jacob, "Convex recovery of continuous domain piecewise constant images from nonuniform fourier samples," *IEEE Transactions on Signal Processing*, vol. 66, no. 1, pp. 236–250, 2017.

[61] G. Ongie and M. Jacob, "Recovery of piecewise smooth images from few fourier samples," in *Sampling Theory and Applications (SampTA), 2015 International Conference on*. IEEE, 2015, pp. 543–547.

[62] ——, "Off-the-grid recovery of piecewise constant images from few fourier samples," *SIAM Journal on Imaging Sciences*, vol. 9, no. 3, pp. 1004–1041, 2016.

[63] G. Ongie, R. Willett, R. D. Nowak, and L. Balzano, "Algebraic variety models for high-rank matrix completion," in *Proceedings of the 34th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, D. Precup and Y. W. Teh, Eds., vol. 70. International Convention Centre, Sydney, Australia: PMLR, 06–11 Aug 2017, pp. 2691–2700. [Online]. Available: http://proceedings.mlr.press/v70/ongie17a.html

[64] G. Ongie and M. Jacob, "A fast algorithm for convolutional structured low-rank matrix recovery," *IEEE Transactions on Computational Imaging*, 2017.

[65] A. Ortega, P. Frossard, J. Kovačević, J. M. Moura, and P. Vandergheynst, "Graph signal processing," *arXiv preprint arXiv:1712.00468*, 2017.

[66] W. Pan, X. Shen, and B. Liu, "Cluster analysis: unsupervised learning via supervised learning with a non-convex penalty." *Journal of Machine Learning Research*, vol. 14, no. 1, pp. 1865–1889, 2013.

[67] H. Pedersen, S. Kozerke, S. Ringgaard, K. Nehrke, and W. Y. Kim, "k-t PCA: Temporally constrained k-t BLAST reconstruction using principal component analysis," *Magnetic resonance in medicine*, vol. 62, no. 3, pp. 706–716, 2009.

[68] S. Poddar and M. Jacob, "Dynamic mri using smoothness regularization on manifolds (storm)," *IEEE Tran. Medical Imaging*, vol. 35, no. 4, pp. 1106–1115, April 2016.

[69] ——, "Low rank recovery with manifold smoothness prior: theory and application to accelerated dynamic MRI," in *International Symposium on Biomedical Imaging (ISBI) 2015*.

[70] ——, "Clustering of data with missing entries using non-convex fusion penalties," *CoRR*, vol. abs/1709.01870, 2017. [Online]. Available: http://arxiv.org/abs/1709.01870

[71] ——, "Recovery of noisy points on band-limited surfaces: Kernel methods re-explained," *arXiv preprint arXiv:1801.00890*, 2018.

[72] S. Poddar, S. G. Lingala, and M. Jacob, "Joint recovery of under sampled signals on a manifold: Application to free breathing cardiac MRI," in *2014 IEEE International Conference on Acoustics, speech and signal processing (ICASSP)*, 2014, pp. 6904–6908.

[73] S. Poddar, Y. Q. Mohsin, D. Ansah, B. Thattaliyath, R. Ashwath, and M. Jacob, "Free-breathing cardiac MRI using bandlimited manifold modelling," *CoRR*, vol. abs/1802.08909, 2018. [Online]. Available: http://arxiv.org/abs/1802.08909

[74] K. P. Pruessmann, M. Weiger, P. Börnert, and P. Boesiger, "Advances in sensitivity encoding with arbitrary k-space trajectories," *Magnetic Resonance in Medicine*, vol. 46, no. 4, pp. 638–651, 2001. [Online]. Available: http://dx.doi.org/10.1002/mrm.1241

[75] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, 2000.

[76] M. Sarkar and T.-Y. Leong, "Fuzzy k-means clustering with missing values." in *Proceedings of the AMIA Symposium*. American Medical Informatics Association, 2001, p. 588.

[77] A. Saxena, M. Prasad, A. Gupta, N. Bharill, O. P. Patel, A. Tiwari, M. J. Er, W. Ding, and C.-T. Lin, "A review of clustering techniques and developments," *Neurocomputing*, 2017.

[78] G. Schiebinger, E. Robeva, and B. Recht, "Superresolution without separation," in *Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP), 2015 IEEE 6th International Workshop on*. IEEE, 2015, pp. 45–48.

[79] B. Schölkopf and A. J. Smola, *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2002.

[80] B. Sharif and Y. Bresler, "Physiologically improved NCAT phantom (PINCAT) enables in-silico study of the effects of beat-to-beat variability on cardiac MR," in *Proceedings of the Annual Meeting of ISMRM, Berlin*, vol. 3418, 2007.

[81] P. J. Shin, P. E. Larson, M. A. Ohliger, M. Elad, J. M. Pauly, D. B. Vigneron, and M. Lustig, "Calibrationless parallel imaging reconstruction based on structured low-rank matrix completion," *Magnetic resonance in medicine*, vol. 72, no. 4, pp. 959–970, 2014.

[82] D. I. Shuman, S. K. Narang, P. Frossard, A. Ortega, and P. Vandergheynst, "The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains," *IEEE Signal Processing Magazine*, vol. 30, no. 3, pp. 83–98, 2013.

[83] A. Singer, "From graph to manifold Laplacian: The convergence rate," *Applied and Computational Harmonic Analysis*, vol. 21, no. 1, pp. 128–134, 2006.

[84] A. Singer, R. Erban, I. G. Kevrekidis, and R. R. Coifman, "Detecting intrinsic slow variables in stochastic dynamical systems by anisotropic diffusion maps," *Proceedings of the National Academy of Sciences*, vol. 106, no. 38, pp. 16 090–16 095, 2009.

[85] A. J. Smola and R. Kondor, "Kernels and regularization on graphs," in *Learning theory and kernel machines*. Springer, 2003, pp. 144–158.

[86] J. B. Tenenbaum, V. d. Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319–2323, 2000.

[87] R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight, "Sparsity and smoothness via the fused lasso," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 67, no. 1, pp. 91–108, 2005.

[88] M. Usman, D. Atkinson, C. Kolbitsch, T. Schaeffter, and C. Prieto, "Manifold learning based ecg-free free-breathing cardiac cine mri," *Journal of Magnetic Resonance Imaging*, vol. 41, no. 6, pp. 1521–1527, 2015.

[89] L. J. van der Maaten, E. O. Postma, and H. J. van den Herik, "Dimensionality reduction: A comparative review," *Journal of Machine Learning Research*, vol. 10, no. 1-41, pp. 66–71, 2009.

[90] K. L. Wagstaff and V. G. Laidler, "Making the most of missing values: Object clustering with partial data in astronomy," in *Astronomical Data Analysis Software and Systems XIV*, vol. 347, 2005, p. 172.

[91] D. O. Walsh, A. F. Gmitro, and M. W. Marcellin, "Adaptive reconstruction of phased array MR imagery," *Magnetic Resonance in Medicine*, vol. 43, no. 5, pp. 682–690, 2000.

[92] J. H. Ward Jr, "Hierarchical grouping to optimize an objective function," *Journal of the American statistical association*, vol. 58, no. 301, pp. 236–244, 1963.

[93] Z. Yang and M. Jacob, "Nonlocal regularization of inverse problems: A unified variational framework," *IEEE Transactions on Image Processing*, vol. 22, no. 8, pp. 3192–3203, Aug 2013.

[94] B. Zhao, J. P. Haldar, and Z.-P. Liang, "PSF Model-Based Reconstruction with Sparsity Constraint: Algorithm and Application to Real-Time Cardiac MRI," *Annual International Conference of the IEEE Engineering in Medicine and Biology Society.*, vol. 2010, pp. 3390–3393, 2010.

[95] C. Zhu, H. Xu, C. Leng, and S. Yan, "Convex optimization procedure for clustering: Theoretical revisit," in *Advances in Neural Information Processing Systems*, 2014, pp. 1619–1627.