A System Centric View of
Modern Structured and Sparse Inference Tasks

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Swayambhoo Jain

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Professor Jarvis Haupt, Advisor

June, 2017

© Swayambhoo Jain 2017
ALL RIGHTS RESERVED

Acknowledgements

First and foremost, I would like to express my deepest gratitude to my doctoral advisor Prof. Jarvis Haupt for his invaluable guidance. Under his mentor-ship I have learned a lot and his constant support and encouragement helped me follow my dreams and aspirations of being able to contribute to the scientific community in my field of interest. He has been a great source of inspiration and someone I look up to and try to emulate when it comes to pursuing a career in research. His helpful nature and kind behavior has helped me *weather* all kinds of obstacles I faced as a doctoral student at University of Minnesota.

I thank Prof. Georgios B. Giannakis, Prof. Nikos Sidiropoulos, and Prof. Arindam Banerjee for agreeing to serve on my committee.

I would also like to thank my professors who put in their heart and soul in teaching us and sincerely sharing with us their knowledge and expertise in various subjects, guiding us, providing useful feedback and helping us get the basics right. I would especially like to thank Prof. Georgios B. Giannakis, Prof. Seung Jun Kim, Prof. Yousef Saad, Prof. Tom Luo, Prof. Keshab Parhi, Prof. Nikos Sidiropoulos, Prof. Arindam Banerjee and Prof. Jarvis Haupt whose teachings and guidance have helped me develop a good understanding of the basics required for research in various topics in Electrical and Computer Engineering (ECE) and for their exclusive feedback on my projects, research works, presentations and during my coursework. I have also had the honor and pleasure of collaborating with them on various projects which resulted in several research ideas, publications and helped me understand my strengths and weaknesses and eventually carve my own niche.

I would also like to thank the funding agencies National Science Foundation (NSF) and Defense Advanced Research Projects Agency (DARPA) who provided financial support for my research activities, and especially the Government of India for providing support for students through their programs for subsidized high quality education.

My experience as a doctoral student of the ECE department of University of Minnesota has been amazing not just because of all of the above mentioned people but very importantly because of the facilities provided by the university and the department to its students. I thank the University and the Department for creating the most comfortable environment for learning.

I would also like to thank Prof. Stefano Gonella and Jeffry Druce with whom I have had opportunity to collaborate on problems pertinent to the field of Civil Engineering which has helped gain experience in inter-disciplinary research.

I would like extend my thanks to Dr. Brian Eriksson and Prof. Kevin S. Xu (now at University of Toledo) for being my mentors during summer internship at Technicolor Research Artificial Intelligence Lab in year 2015, and who continue to provide support and guidance. I also want to thank my colleagues during this internship, especially Urvashi Oswal, who has become an active research collaborator.

I also express my sincere gratitude to my peers Akshay Soni, Mojtaba Kadkhodaie, Alexander Gutierrez, Sirisha Rambhatla, Xingguo Li, Abhinav V. Sambasivan, Di Xiao, and Scott Sievert for their patience and support. Having an over-enthusiastic labmate like me could be a challenge sometimes, especially coping up with the distractions I caused in form of unsolicited ideas, opinions and advice. I feel honored to have worked with my peers who are experts in their respective domains and have always obliged me by answering my questions, providing feedback and collaborating on various projects and research topics. I would also like to thank Nitin Jain, Balasubramanian Gopalakrishnan, Aritra Konar, Meisam Razaviyayn, Morteza Mardani, Kejun Huang, Yu Zhang, Brian Baingana, and my other peers across the labs in ECE department for providing an alternate perspective and sharing their learning and research experiences.

But the journey to University of Minnesota would not have been possible without the guidance and teachings of professors who guided me through my undergraduate studies at National Institute of Technology - Calicut (NITC). I would like to especially thank Prof. P. P. Deepthi, Prof. A. V. Babu and Prof. G. Abhilash at NITC for setting the examples of how pursuing one's field of interest is a rewarding experience. I would also like to thank Dr. Neelesh B. Mehta at Indian Institute of Science Bangalore whose guidance and encouragement helped me transition from industry back to academics.

Last but not the least I would like to thank my family for always being by my side, encouraging me and providing me the required support to pursue my dreams and aspirations. They never complained even when I forgot a birthday due to an impending deadline or couldn't make it home for an important occasion. The glow on their faces when I discuss my successes and the worry followed by the smile of encouragement when I discuss my problems is priceless and I hope I make them proud and am able to make it only easy for them going forward by matching the support and encouragement I have had the privilege of.

Swayambhoo Jain, Minneapolis, MN, May 2017.

Abstract

We are living in the era of *data deluge* wherein we are collecting unprecedented amount of data from variety of sources. Modern inference tasks are centered around exploiting structure and sparsity in the data to extract relevant information. This thesis takes an end-to-end system centric view of these inference tasks which mainly consist of two sub-parts (i) data acquisition and (ii) data processing. In context of the data acquisition part of the system, we address issues pertaining to noise, clutter (the unwanted extraneous signals which accompany the desired signal), quantization, and missing observations. In the data processing part of the system we investigate the problems that arise in resource-constrained scenarios such as limited computational power and limited battery life.

The first part of this thesis is centered around computationally-efficient approximations of a given linear dimensionality reduction (LDR) operator. In particular, we explore the partial circulant matrix (a matrix whose rows are related by circular shifts) based approximations as they allow for computationally-efficient implementations. We present several theoretical results that provide insight into existence of such approximations. We also propose a data-driven approach to numerically obtain such approximations and demonstrate the utility on real-life data.

The second part of this thesis is focused around the issues of noise, missing observations, and quantization arising in matrix and tensor data. In particular, we propose a sparsity regularized maximum likelihood approach to completion of matrices following sparse factor models (matrices which can be expressed as a product of two matrices one of which is sparse). We provide general theoretical error bounds for the proposed approach which can be instantiated for variety of noise distributions. We also consider the problem of tensor completion and extend the results of matrix completion to the tensor setting. The problem of matrix completion from quantized and noisy observations is also investigated in as general terms as possible. We propose a constrained maximum likelihood approach to quantized matrix completion, provide probabilistic error bounds for this approach, and numerical algorithms which are used to provide numerical evidence for the proposed error bounds.

The final part of this thesis is focused on issues related to clutter and limited battery life in signal acquisition. Specifically, we investigate the problem of compressive measurement design under a given sensing energy budget for estimating structured signals in structured clutter. We propose a novel approach that leverages the prior information about signal and clutter to judiciously allocate sensing energy to the compressive measurements. We also investigate the problem of processing Electrodermal Activity (EDA) signals recorded as the conductance over a user's skin. EDA signals contain information about the user's neuron firing and psychological state. These signals contain the desired information carrying signal superimposed with unwanted components which may be considered as clutter. We propose a novel compressed sensing based approach with provable error guarantees for processing EDA signals to extract relevant information, and demonstrate its efficacy, as compared to existing techniques, via numerical experiments.

Contents

Acknowledgements **i**

Abstract **iv**

List of Tables **xi**

List of Figures **xii**

1 Introduction **1**

1.1 Computationally-efficient approximations to arbitrary linear dimensionality reduction operators . . . 2

1.2 Noisy matrix and tensor completion under sparse factor models . . . 3

1.3 Matrix completion from noisy and quantized observations . . . 4

1.4 Compressive measurement designs for estimating structured signals in structured clutter . . . 4

1.5 A compressed sensing based decomposition of electrodermal activity signals 5

1.6 Published results . . . 5

2 Computationally-efficient approximations to arbitrary linear dimensionality reduction operators **7**

2.1 Problem statement and our contributions . . . 8

2.2 Connections to existing work . . . 9

2.3 Preliminaries and notations . . . 10

2.4 Understanding the fundamental nature of the problem . . . 11

2.5 A fundamental (negative) approximation result for partial circulant matrices . . . 13
2.6 Approximation with partial circulant matrices with post-processing . . . 14
2.7 The data driven approach . . . 17
2.7.1 Data driven approach for partial circulant matrices . . . 17
2.7.2 Data driven approach for partial circulant matrices with post-processing . . . 18
2.8 Algorithms for the data driven approach . . . 19
2.8.1 Data driven approach for partial circulant matrices . . . 19
2.8.2 Data driven approach for partial circulant matrices with post-processing . . . 22
2.9 Experimental evaluation . . . 25
2.10 Summary . . . 27
2.11 Appendix . . . 27
2.11.1 Proof of lemma 2.4.1 . . . 27
2.11.2 Proof of theorem 2.5.1 . . . 28
2.11.3 Proof of lemma 2.6.1 . . . 30
2.11.4 Proof of theorem 2.6.1 . . . 32
2.11.5 Proof of lemma 2.6.2 . . . 33
2.11.6 Proof of lemma 2.7.1 . . . 34
2.11.7 Proof of lemma 2.11.2 . . . 35
2.11.8 Proof of lemma 2.11.3 . . . 36
2.11.9 Proof of theorem 2.8.1 . . . 38

3 Noisy matrix and tensor completion under sparse factor models 39
3.1 Preliminaries and notations . . . 40
3.2 Noisy matrix completion for sparse factor models . . . 41
3.3 Noisy tensor completion for tensors with a sparse canonical polyadic factor 43
3.3.1 Data model . . . 44
3.3.2 Observation setup . . . 44
3.3.3 Estimation procedure . . . 45
3.3.4 General error bound . . . 46

3.3.5 Implication for Gaussian noise 47
3.3.6 The algorithmic framework 48
3.3.7 Experimental evaluation 50
3.4 Summary 52
3.5 Appendix 52
3.5.1 Proof of theorem 3.3.1 52
3.5.2 Proof of lemma 3.5.1 55
3.5.3 Proof of corollary 3.3.1 58

4 Matrix completion from noisy and quantized observations 61
4.1 Preliminaries and notations 62
4.2 Observation model 63
4.3 Estimation procedure 63
4.4 Towards error guarantees 65
4.5 Error guarantees based on covering number approach 65
4.5.1 Set of low rank matrix with bounded entries 68
4.5.2 Sparse factor model with bounded entries 69
4.6 Error guarantees based on the chaining principle 70
4.6.1 Set of low rank matrices 73
4.7 Further discussion on the error bounds 74
4.8 Numerical algorithms 74
4.8.1 For convex $\mathcal{X}$ 75
4.8.2 For non-convex $\mathcal{X}$ 75
4.9 Experimental evaluation 78
4.10 Summary 80
4.11 Appendix 80
4.11.1 Useful lemmata 80
4.11.2 Proof of lemma 4.5.1 89
4.11.3 Proof of lemma 4.5.2 90
4.12 Proof of lemma 4.6.1 91
4.13 Proof of lemma 4.6.2 92
4.14 Proof of corollary 4.6.1 94

5 Compressive measurement designs for estimating structured signals in structured clutter 96

5.1 Problem statement and our contributions . . . 97

5.2 Connections with existing work . . . 98

5.3 Quantifying prior information . . . 99

5.4 Choosing the set of sensing matrices $\mathcal{A}$ . . . 100

5.5 Choosing set of estimators $\mathcal{X}$ . . . 100

5.6 Our proposed design approach . . . 101

5.6.1 Solving **P0** . . . 103

5.6.2 Solving **P1** . . . 104

5.7 Experimental evaluation . . . 105

5.8 Summary . . . 107

6 A compressed sensing based decomposition of electrodermal activity signals 108

6.1 Related work . . . 110

6.2 Model . . . 111

6.2.1 Model definition . . . 113

6.2.2 SCR events signal model . . . 114

6.2.3 Baseline model . . . 115

6.2.4 Bounded noise model . . . 116

6.2.5 Problem overview . . . 116

6.3 EDA signal decomposition . . . 116

6.3.1 Dealing with the baseline signal . . . 117

6.3.2 EDA signal preprocessing . . . 118

6.3.3 Error guarantees . . . 119

6.4 Experiments . . . 121

6.4.1 Synthetic data experiment . . . 122

6.4.2 Experiments with real-world EDA data . . . 124

6.5 Summary . . . 127

6.6 Appendix . . . 128

7 Future Directions **130**

7.1 Computationally-efficient approximations to arbitrary linear dimensionality reduction operators 130

7.2 Noisy matrix and tensor completion under sparse factor models. 131

7.3 Matrix completion from noisy and quantized observations 132

References **133**

List of Tables

5.1 Iterative algorithm for solving the sensing matrix design problem (5.4). 103

6.1 EDA Signal Notation Summary 114

6.2 AUC values for SCR event detection at multiple sampling rates for various approaches on real data experiment. 127

List of Figures

1.1 Data processing pipeline 1

2.1 A representative figure showing the geometric interpretation of the problem after vectorization. 12

2.2 Geometric interpretation of *Rubik's* score. 13

2.3 A representative figure showing the geometric implication of "post-processing" matrix in terms of denser union of subspaces providing better approximation. 16

2.4 Data-driven approximation of matrix comprising of top 300 principal components of a training set of images from COIL-20 database. The first panel (left to right) contains log average relative approximation error vs. m' with various circulant approximations to the given matrix. The second panel plots shows the ratio of average time taken by these circulant approximation and $\mathbf{A}$ vs. the log average relative approximation error. The third panel contains the histograms of errors by the circulant approximation matrix obtained by Algorithm 3. 25

3.1 Plot for log per-entry approximation error vs the log sampling rate for the two ranks $F = 5, 15$. The slope at the higher sampling rates is approximately -1 (the rate predicted by our theory) in both cases. . . 52

4.1 Observation model under quantized setting 63

4.2 Results of synthetic experiments for quantized matrix completion under sparse factor models. The panel (a) contains log per-entry squared error $\log\left(\frac{\|\hat{\mathbf{X}}-\mathbf{X}^*\|_F^2}{n_1 n_2}\right)$ vs. log sampling rate $\log_{10}(\gamma)$ plots for $K = 8, 4, 2$. Panel (b) contains per-entry squared error $\frac{\|\mathbf{X}-\mathbf{X}^*\|_F^2}{n_1 n_2}$ vs. number of symbols in quantizer K plots for $\gamma = 0.10, 0.25$. 79

5.1 Reconstruction SNR $= 20 \log \frac{\|\mathbf{x}\|_2}{\|\mathbf{x}-\hat{\mathbf{x}}\|_2}$ vs. sensing energy budget for several different compressive measurement strategies (see text for details). Panels (a)-(d) correspond to $m = 20, 40, 60, 80$ measurements, respectively. Higher SNR values correspond to better reconstructions. Our proposed approach (blue dotted line, circle markers) outperforms each of the other measurement strategies examined. 105

6.1 An example of EDA signal where the Skin Conductance Responses (SCRs) resulting from SCR events signal are shown. 112

6.2 Observation model showing the various components in the observed EDA signal. 113

6.3 An example EDA signal collected using a commercially available wearable EDA sensor showing the impact of baseline shifts due to movement. . . 115

6.4 Block diagram showing the SCR events signal recovery using compressed sensing based decomposition. 118

6.6 The impulse response $\mathbf{h}$ was obtained by sampling the function $f(u) = 2\left(e^{-\frac{u}{\tau_1}} - e^{-\frac{u}{\tau_2}}\right)$ for $u \geq 0$ and $f(u) = 0$ otherwise. Here $\tau_1 = 10, \tau_2 = 1$ and the is function sampled at the rate of 4 samples per second in the interval $u \in [0, 40]$. 122

6.8 ROC curves for real data SCR event detection experiment at the sampling rate of 4 Hz. Our compressed sensing-based approaches is compared with a variation of our approach with positivity constraints, the non-negative deconvolution approach in Ledalab, the cvxEDA convex optimization based approach, and the raw EDA signal. 126

Chapter 1

Introduction

We are living in an exciting era of *"data-deluge"* wherein we are collecting an unprecedented amount of data. This *"data-deluge"* has been fueled by increasing hardware capability and efficient software front-ends and back-ends. In last decade and a half we have witnessed wide variety of practical problems being solved from data-centric point of view. Terms like *Data Science* and *Big Data* have become colloquial. We are witness to lots of new exciting developments in these fields. However, the typical data processing pipeline has more or less remained the same. Abstractly, the data processing pipeline (shown in figure 1.1) consists of two main modules:

- *Data acquisition* module that acquires data from various sources such as natural images, speech, biometric signals etc.
- *Data processing* module that processes the acquired data to produce the desired output for the given inference task.

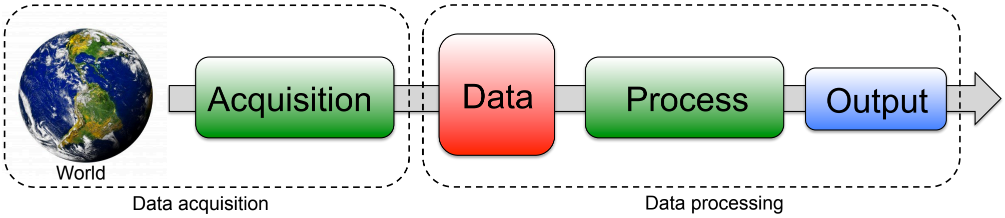


Figure 1.1: Data processing pipeline

Many tasks in signal processing and machine learning can be broadly captured by the abstract data processing pipeline shown in figure 1.1. As we transition to an era of *Internet of Things* (IoT) in which signals from various sources will be used to produce information by using extremely low cost devices, addressing issues related to data acquisition and efficient data processing in resource constrained environments will become crucial. Modern inference tasks are centered around exploiting structure and sparsity in the data to extract relevant information. This thesis is centered around various problems which deal with practical issues arising in these inference tasks. The topics of research discussed in this thesis lay specific emphasis on modeling the data acquisition process explicitly, and improving/modifying the classical data processing techniques under the given resource constraints arising due to limited battery life, computational power, etc. Below we summarize the main research topics covered in this thesis.

1.1 Computationally-efficient approximations to arbitrary linear dimensionality reduction operators

The amount of data being acquired is ever increasing in volume and dimensionality. This calls for efficient implementations of even very simple data processing tasks such as linear dimensionality reduction (LDR). The LDR operator can be represented as a matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$, with $m < n$. Operating with $\mathbf{A}$ on an arbitrary vector $\mathbf{x} \in \mathbb{R}^n$ generally requires $\mathcal{O}(mn)$ operations, which can be superlinear in n for even modest values of m (e.g., when $m = n^\beta$ for $\beta \in (0, 1]$ the complexity is $\mathcal{O}(n^{1+\beta})$). This topic examines the problem of approximating a given matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ by a partial circulant matrix whose rows are related by circular shifts. Our investigation is motivated by the implementability benefits of these structured approximations, both in hardware (as sampled outputs of certain linear time invariant systems) and software (via fast Fourier transform based implementations). Our contributions on this topic are provided in chapter 2 where we show (analytically) that while most large LDR matrices cannot themselves be well-approximated (in a Frobenius sense) by partial circulant matrices, a slightly generalized framework that allows for modest linear post-processing does enable accurate partial circulant based approximations of general LDR matrices over a restricted domain of inputs. We also propose algorithms based on alternating minimization and sparse matrix factorization for identifying the factors comprising the approximations, and provide experimental evidence to demonstrate the potential efficacy of this framework.

1.2 Noisy matrix and tensor completion under sparse factor models

Often in practice the data acquisition process suffers from noisy and missing observations which usually arise due to hardware resource constraints or sometimes they are inherent in the observation setup itself. In the first part of this topic we examine the task of noisy matrix completion that involves estimating the matrix from noisy observations collected at a subset of its entries. Our specific focus is on the set of matrices following *sparse factor models* – the matrices which may be expressed as a product of two matrices, one of which is sparse. Such matrices arise in wide variety of applications including subspace clustering and dictionary learning. Leveraging the structure, we propose sparsity-regularized maximum likelihood estimation for matrix completion. Our main contribution comes in the form of generic estimation error bounds for sparsity-regularized maximum likelihood estimators. The bounds are general enough to be instantiated for a variety of noise distributions.

In the second part of this topic we extend the results of matrix completion to the tensor setting. Specifically, we focus on 3-way tensors that admit *sparse CP decomposition* by which we mean that one of the canonical polyadic or CANDECOMP/PARAFAC (CP)-factors is sparse. Tensors admitting such structure arise in many applications involving electroencephalography (EEG) data, neuroimaging using functional magnetic resonance imaging (MRI), and many others. We consider sparsity-regularized maximum likelihood estimation approach for tensor completion and provide generic estimation error guarantees. While the task of tensor completion can be posed as a matrix completion problem by stacking the slices of the tensor, we demonstrate the specific advantage of considering the tensor structure. In particular, we instantiate the generic error bounds for the Gaussian noise case and show that tighter error bounds can be obtained if tensor structure is considered. We also provide an alternating direction method of multiplier-type algorithm for approximately solving the complexity-regularized maximum likelihood problem for tensor completion tasks and provide experimental evidence for the error analyses. In chapter 3 we discuss the aforementioned contributions on this topic in detail.

1.3 Matrix completion from noisy and quantized observations

In this topic, we consider the general problem of matrix estimation from noisy and quantized measurements taken at a subset of its entries. Specifically, we assume that the matrix to be completed lies in a structured set and consider the constrained maximum likelihood estimates. We provide two types of probabilistic estimation error bounds obtained via covering number based approach and a more involved chaining argument based approach. We instantiate these bounds for the set of low rank matrices as well as matrices following sparse factor models. For the completion of matrices following sparse factor model we propose an alternating direction method of multiplier-type algorithm for approximately solving the constrained maximum likelihood problem, and provide empirical evidence for the theoretical bounds. In chapter 4 we discuss our contributions on this topic in detail.

1.4 Compressive measurement designs for estimating structured signals in structured clutter

In many practical applications the acquired data is corrupted with pre-measurement clutter and post-measurement noise. Typically, in such applications practical constraints arise due to limited resources. However, given prior knowledge about the signal, clutter and noise, it is possible to improve the data acquisition process under the given practical constraints. In this topic we provide systematic investigation of leveraging prior knowledge under finite sensing energy constraint for estimating structured signal in structured clutter and propose a method for designing a compressive measurement strategy under the given sensing energy budget. Experimental results on synthetic data demonstrate that the proposed approach outperforms traditional random compressive measurement designs, which are agnostic to the prior information, as well as several other knowledge-enhanced sensing matrix designs based on more heuristic notions. Chapter 5 provides detailed discussion of our work on this topic.

1.5 A compressed sensing based decomposition of electrodermal activity signals

Electrodermal Activity, or EDA, is typically recorded as the conductance over a person's skin, near concentrations of sweat glands (*e.g.,* palm of the hand or finger tips). EDA signals have been shown to include significant information pertaining to human neuron firing and psychological arousal. The measurement and analysis of EDA signals is critical to a wide variety of applications ranging from health analytics to market research. The central challenge in analysis of EDA signals is that it is superimposed with numerous extraneous components, collectively termed as the baseline signal. This is an instance of a scenario where the desired information bearing signal is corrupted with undesired non-informative clutter. We propose a novel approach to process the EDA signals, which involves a simple pre-processing followed by compressed sensing based decomposition. Our approach alleviates the effects of baseline with provable bounds on the recovery of user responses. Through numerical experiments on synthetic as well as real-world EDA signals from wearable sensors, we demonstrate that our approach leads to more accurate recovery of user responses as compared to the existing techniques. Chapter 6 provides detailed discussion of our work on this topic.

1.6 Published results

The topics investigated in this Ph.D. thesis have resulted in several publications. In particular, the abridged work on the topic of computationally-efficient approximations to arbitrary linear dimensionality reduction operators was published as conference paper [1] and the full version is in preparation to be submitted to the Institute of Electrical and Electronic Engineers (IEEE) Transactions on Signal Processing. The investigations related to matrix/tensor completion for sparse factor models have resulted in one journal published in IEEE Transactions on Information Theory [2], and two conference publications [3, 4]. The research into the topic of compressive measurement designs for estimating structured signals in structured clutter has resulted in two conference publications [5, 6]. Our investigations into compressed sensing based decomposition of electrodermal activity signals has resulted in one journal publication in the IEEE Transactions on Biomedical Engineering [7]. Other publications related to Ph.D. work reported in this thesis but not discussed in detail consists of two journal publications [8,9] and four conference publications [10–13].

Chapter 2

Computationally-efficient approximations to arbitrary linear dimensionality reduction operators

Numerous tasks in signal processing, statistics, and machine learning employ dimensionality reduction methods to facilitate the processing, visualization, and analysis of (ostensibly) high-dimensional data in (more tractable) low-dimensional spaces. [1] Among the myriad of dimensionality reduction techniques in the literature, *linear dimensionality reduction* (LDR) methods remain among the most popular and widely-used.

One well-known example is principal component analysis [14] characterized by a linear dimensionality reduction (LDR) operator designed to maximally preserve the variance of the original data in the projected space, and which has been widely used for data compression, denoising, and as a dimensionality reducing pre-processing step for other analyses (e.g., clustering, classification, etc.) [15]. Other classical data analysis methods that employ specialized LDR operators include linear discriminant analysis (where the operator is designed to preserve separations among original data points

[1] Some of the results reported in this chapter have been published without proof in [1]. The complete proofs of these results are provided in this chapter.

belonging to disparate classes), and canonical correlations analysis (where the operator maximizes correlations among projected data points). See the survey paper [16] for many additional examples.

In recent years, LDR methods have also been utilized for universal "precompression" in high-dimensional inference tasks. For example, fully random LDR operators are at the essence of the initial investigations into *compressed sensing* (CS) (see, e.g., [17]); and other, more structured, LDR operators – both non-adaptive (see, e.g., [6, 18–26]) and adaptive (see, e.g., [27–43]) – have also been examined recently in the context of CS and sparse inference. The computational efficiency of LDR methods is often cited as one of their primary virtues; however, as data of interest become increasingly large-scale (high-dimensional, and numerous), even the relatively low computational complexity associated with LDR methods can become significant. Here we investigate the utility of employing *partial circulant approximations* to general LDR operators; partial circulant matrices admit fast implementations (via convolution or Fourier transform methods, and subsampling), and their use as surrogates to arbitrary LDR matrices may provide significant computational efficiency improvements in practice.

2.1 Problem statement and our contributions

We represent an arbitrary (here, real) linear dimensionality reduction (LDR) operator as a matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$, with $m < n$. Operating with $\mathbf{A}$ on an arbitrary vector $\mathbf{x} \in \mathbb{R}^n$ generally requires $\mathcal{O}(mn)$ operations, which can be superlinear in n for even modest values m (e.g., when $m = n^\beta$ for $\beta \in (0, 1]$ the complexity is $\mathcal{O}(n^{1+\beta})$). Here, we seek computationally-efficient approximations of $\mathbf{A}$ implementable via convolution and downsampling (and, perhaps, modest post-processing).

We first consider approximating $\mathbf{A}$ as $\mathbf{A} \approx \mathbf{SC}$, where $\mathbf{SC}$ is the *partial circulant* matrix obtained by choosing m distinct rows from an $n \times n$ circulant matrix $\mathbf{C}$ using row-subsampling matrix $\mathbf{S}$ containing m rows that are a (permuted) subset of rows of $\mathbf{I}_n$ (the $n \times n$ identity). These *partial circulant* approximations enjoy an implementation complexity of $\mathcal{O}(n \log n)$ (owing to fast Fourier transforms) and storage cost of $\mathcal{O}(m + n)$. Despite these potential benefits, our first result here is negative – we establish that "most" LDR matrices are not well-approximated (in a Frobenius sense) by partial circulant matrices.

We then propose a generalization that uses approximations of the form $\mathbf{A} \approx \mathbf{PSC}$, where $\mathbf{C}$ and $\mathbf{S}$ are as above, except that $\mathbf{S}$ has some $m' \geq m$ rows, and $\mathbf{P}$ is an $m \times m'$ "post-processing" matrix. Operating with such matrices requires $\mathcal{O}(mm' + n \log n)$ operations in general, and can be as low as $\mathcal{O}(n \log n)$, e.g., when $m' = \mathcal{O}(n^{1/2})$. Though the storage cost slightly increases to $\mathcal{O}(mm' + m' + n)$ it is still better than $\mathcal{O}(mn)$. In addition to these, the convolutional nature of these approximations makes them viable for implementation using LTI systems, or for applications where such models arise naturally (e.g., RADAR). Within this framework, we exploit the fact that signals of interest often reside on a *restricted input domain* (e.g., a union of subspaces, manifold, etc.), so we may restrict our approximation to mimicking the action of $\mathbf{A}$ on these inputs. We provide a concise argument establishing the efficacy of this more general approach for certain restricted inputs, describe a *data-driven* approach to learning the factors of the approximating matrix, and provide empirical evidence to demonstrate the viability of this approach.

2.2 Connections to existing work

Circulant approximations to square matrices are classical in linear algebra; for example, circulant preconditioners for linear systems were examined in [44–46]. In a line of work motivated by "optical information processing," several efforts have examined fundamental aspects of approximating square matrices by products of circulant and diagonal matrices [47–51]. Here, our focus is on LDR matrices (not square matrices), so results from these works are not directly applicable here. Partial circulant matrices have been used in signal processing and high dimensional data analysis. Many random constructions of partial circulant matrices can be shown to satisfy the restricted isometry condition for sparse signals and therefore they have been used in compressive sensing [52–55]. In high dimensional data analysis they are used as computationally efficient alternative for stably embedding high dimensional data vectors into lower dimensions [56–58]. In these works too the partial circulant matrices are constructed randomly. In contrast to these works, here our aim is to approximate the action of *given* LDR matrix, not necessarily to perform JL embeddings.

2.3 Preliminaries and notations

We introduce some preliminary concepts and notations that will be used throughout this chapter. For $\mathbf{A} \in \mathbb{R}^{m\times n}$, we denote its m individual rows by $\mathbf{A}_{i,:}$ for $i = 1, \dots, m$ and its n columns by $\mathbf{A}_{:,j} \in \mathbb{R}^m$ for $j = 1, \dots, n$. The *row-wise* vectorization $\mathbf{A}$ is denoted by $\text{vec}(\mathbf{A}) = [\mathbf{A}_{1,:}, \cdots, \mathbf{A}_{m,:}]^m$. The squared Frobenius norm of $\mathbf{A}$ is $\|\mathbf{A}\|_F^2 = \sum_{i,j} |A_{i,j}|^2$, and $\|\mathbf{A}\|_{2,1} = \sum_{j=1}^n \|\mathbf{A}_{:,j}\|_2$, where $\|\mathbf{A}_{:,j}\|_2$ is the Euclidean norm of $\mathbf{A}_{:,j}$. Finally, $\|\mathbf{A}\|_{2\to 2} \triangleq \sup_{\mathbf{x}\neq \mathbf{0}} \|\mathbf{Ax}\|_2/\|\mathbf{x}\|_2$ denotes the spectral norm of $\mathbf{A}$. A $n \times n$ identity matrix is denoted by $\mathbf{I}_n$ and set obtained by choosing some $m < n$ distinct rows of identity matrix $\mathbf{I}_n$ is denoted by $\mathcal{S}_m$.

For analytical description of a circulant matrix we use the notion of "right rotation" matrix defined as

$$\mathbf{R} = \begin{bmatrix} \mathbf{0}_{(n-1)\times 1} & \mathbf{I}_{n-1} \\ 1 & \mathbf{0}_{1\times(n-1)} \end{bmatrix}.$$

The post-multiplication of any row vector $\mathbf{c} \in \mathbb{R}^n$ by $\mathbf{R}$ gives a circular shift to the right by one position as follows $\mathbf{c}^T\mathbf{R} = [c_n, c_1, \cdots, c_{n-1}]$. Analogously, post-multiplying a row vector by $\mathbf{L} = \mathbf{R}^T$ implements a circular shift to the left by one position; note that $\mathbf{LR} = \mathbf{I}_n$. We represent an $n \times n$ (real) circulant matrix by

$$\mathbf{C} = \begin{bmatrix} c_1 & c_2 & \cdots & c_n \\ c_n & c_1 & \cdots & c_{n-1} \\ & & \ddots & \\ c_2 & c_3 & \cdots & c_1 \end{bmatrix} \tag{2.1}$$

where $\mathbf{c} = [c_1 \ \cdots \ c_n]^T \in \mathbb{R}^n$. We let $\mathcal{C}_n$ denote the set of all (real) n-dimensional circulant matrices of the form (2.1) which can be described in terms of "right rotation matrix" $\mathbf{R}$ as

$$\mathcal{C}_n = \left\{ \begin{bmatrix} \mathbf{c}^T \\ \mathbf{c}^T\mathbf{R} \\ \vdots \\ \mathbf{c}^T\mathbf{R}^{n-1} \end{bmatrix} \middle| \mathbf{c} \in \mathbb{R}^n \right\}. \tag{2.2}$$

All circulant matrices $\mathbf{C} \in \mathcal{C}_n$ can be factorized as follows

$$\mathbf{C} = \mathbf{F}^{-1}\text{diag}(\mathbf{Fe})\mathbf{F}, \tag{2.3}$$

where $F_{jk} = e^{-2jk\pi i/n}$ and $\mathbf{e}$ is first column of $\mathbf{C}$. A *partial* circulant matrix of size $m \times n$ is defined as the matrix which is obtained by sampling some m unique rows from rows of a circulant matrix. We denote the set of $m \times n$ *partial* circulant matrix by $\mathcal{P}C_{m,n}$ analytically defined as follows

$$\mathcal{P}C_{m,n} = \left\{ \begin{bmatrix} \mathbf{c}^T\mathbf{R}^{f_1} \\ \vdots \\ \mathbf{c}^T\mathbf{R}^{f_m} \end{bmatrix} \middle| \, \mathbf{f} \in \mathcal{F}, \mathbf{c} \in \mathbb{R}^n \right\}, \tag{2.4}$$

where $\mathcal{F} = \left\{ \mathbf{f} = [f_1, \cdots, f_m] \in \{0, ..., n-1\}^m \middle| f_i \neq f_j \ \forall i \neq j \right\}$.

2.4 Understanding the fundamental nature of the problem

As above, we let $\mathbf{A} \in \mathbb{R}^{m \times n}$ denote the LDR matrix we aim to approximate using a partial circulant matrix of the same dimensions. Here, we consider a tractable (and somewhat natural) choice of the approximation error metric, and seek to find the matrix $\mathbf{Z} \in \mathcal{P}C_{m,n}$ closest to $\mathbf{A}$ in the Frobenius sense. In this setting, the minimum approximation error is

$$\mathcal{E}_{\mathcal{P}C_{m,n}}(\mathbf{A}) = \min_{\mathbf{Z} \in \mathcal{P}C_{m,n}} \|\mathbf{A} - \mathbf{Z}\|_F^2. \tag{2.5}$$

Evaluating (2.5) is a non-convex optimization problem, owing to the product as well as combinatorial nature of elements of $\mathcal{P}C_{m,n}$. However, an intuitive geometric insight can be obtained if we *row-wise* vectorize the matrices in the problem (2.5) and solve the following equivalent vectorized problem

$$\mathcal{E}_{\mathcal{P}C_{m,n}}(\mathbf{A}) = \min_{\mathbf{Z} \in \mathcal{P}C_{m,n}} \|\mathbf{A} - \mathbf{Z}\|_F^2 = \min_{\mathbf{z} \in \text{vec}(\mathcal{P}C_{m,n})} \|\text{vec}(\mathbf{A}) - \mathbf{z}\|_2^2,$$

where $\text{vec}(\mathcal{P}C_{m,n})$ is the set obtained by *row-wise* vectorizing the matrix set $\mathcal{P}C_{m,n}$. It is easy to see that $\text{vec}(\mathbf{A}) \in \mathbb{R}^{mn}$ whereas owing to the description of set $\mathcal{P}C_{m,n}$ in (2.4) each vector $\mathbf{z} \in \text{vec}(\mathcal{P}C_{m,n})$ has the form $\mathbf{z} = \mathbf{c}^T[\mathbf{R}^{f_1}, \cdots, \mathbf{R}^{f_m}]$ where $\mathbf{c} \in \mathbb{R}^n$ and $\mathbf{f} \in \mathcal{F}$. For a particular $\mathbf{f} \in \mathcal{F}$, the vectorized partial circulant matrices lie in the row-subspace of the matrix $[\mathbf{R}^{f_1}, \cdots, \mathbf{R}^{f_m}]$. As the matrix set $\mathcal{P}C_{m,n}$ is obtained by taking union over the elements in set $\mathcal{F}$ its vectorized version $\text{vec}(\mathcal{P}C_{m,n})$ has union of subspaces in $\mathbb{R}^{mn}$. There are precisely $|\mathcal{F}| = \binom{n}{m} m! \leq n^m$ subspaces. A representative picture for intuitive understanding of this geometric interpretation is shown in figure 2.1.

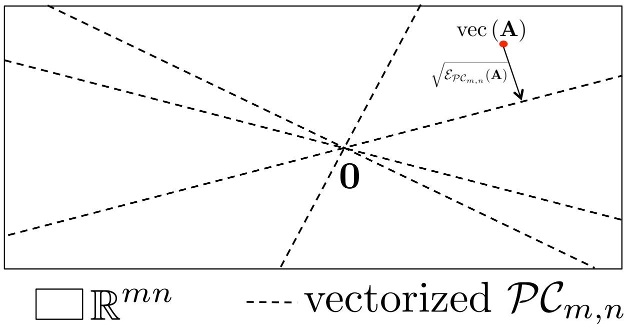

vec (**A**)
$\sqrt{\mathcal{E}_{\mathcal{PC}_{m,n}}(\mathbf{A})}$
0
$\mathbb{R}^{mn}$
vectorized $\mathcal{PC}_{m,n}$

Figure 2.1: A representative figure showing the geometric interpretation of the problem after vectorization.

Thus, how accurately a given LDR matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ can be approximated via a partial circulant matrix in $\mathcal{P}C_{m,n}$ depends on how well the union of subspaces represented by vectorized $\mathcal{P}C_{m,n}$ covers $\mathbb{R}^{mn}$. We obtain a precise characterization of the minimum achievable approximation error for a given $\mathbf{A}$ in the following lemma.

Lemma 2.4.1. For $\mathbf{A} \in \mathbb{R}^{m \times n}$, we have

$$\mathcal{E}_{\mathcal{P}C_{m,n}}(\mathbf{A}) = \|\mathbf{A}\|_F^2 - R^2(\mathbf{A}), \tag{2.6}$$

where $R(\mathbf{A})$ is the *Rubik's Score* of $\mathbf{A}$ defined as

$$R(\mathbf{A}) = \max_{\mathbf{f} \in \mathcal{F}} \frac{\| \sum_{i=1}^{m} \mathbf{A}_{i,:} \mathbf{L}^{f_i} \|_2}{\sqrt{m}}, \tag{2.7}$$

and $\mathcal{F} = \Big\{ \mathbf{f} = [f_1 ... f_m] \in \{0, ..., n-1\}^m \Big| f_i \neq f_j \ \forall i \neq j \Big\}$.

Proof. The proof is outline in appendix section 2.11.1. □

The above lemma reveals the fundamental quantity $R(\mathbf{A})$ in (2.7) which determines how well a given LDR matrix can be approximated by a partial circulant matrix. We call $R(\mathbf{A})$ the *Rubik's Score* of the matrix $\mathbf{A}$, inspired by the fact that $R(\mathbf{A})$ is maximized when the circular shifts of the rows $\{\mathbf{A}_{i,:}\}$ are "maximally aligned". The *Rubik's Score* can be shown to have following properties:

- $R(\mathbf{A}) = \|\hat{\mathbf{A}}\|_F$ where $\hat{\mathbf{A}} = \arg\min_{\mathbf{Z}\in\mathcal{P}C_{m,n}} \ \|\mathbf{A}-\mathbf{Z}\|_F^2$, geometrically it corresponds to the Frobenius norm based distance of best approximation from origin.

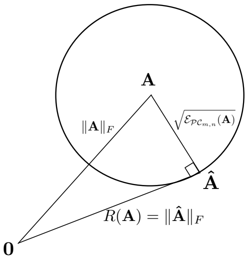

$\mathbf{A}$
$\|\mathbf{A}\|_F$
$\sqrt{\mathcal{E}_{\mathcal{P}\mathcal{C}_{m,n}}(\mathbf{A})}$
$\hat{\mathbf{A}}$
$R(\mathbf{A}) = \|\hat{\mathbf{A}}\|_F$
$\mathbf{0}$

Figure 2.2: Geometric interpretation of *Rubik's* score.

- $0 \leq R(\mathbf{A}) \leq \|\mathbf{A}\|_F$.
- For all $\mathbf{A} \in \mathcal{P}C_{m,n}$, $R(\mathbf{A}) = \|\mathbf{A}\|$ which implies that $\mathcal{E}_{\mathcal{P}C_{m,n}}(\mathbf{A}) = 0$ for all $\mathbf{A} \in \mathcal{P}C_{m,n}$.

2.5 A fundamental (negative) approximation result for partial circulant matrices

The Lemma 2.4.1 gives fundamental insight into approximating the given LDR matrix with a partial circulant matrices. However, it yields limited interpretation of achievable error for any particular $\mathbf{A}$ beyond the cases when the cases when $\mathbf{A}$ is itself a partial circulant matrix. We gain additional insight using a *probabilistic technique* – instead of quantifying the approximation error for a fixed $\mathbf{A}$, we consider instead drawing matrices $\mathbf{A}$ randomly, so that their row spaces are distributed uniformly at random on $\mathrm{Gr}(m, n)$, the Grassmannian manifold of m-dimensional linear subspaces of $\mathbb{R}^n$. We then quantify the *proportion* of matrices so drawn whose (optimal) partial circulant approximation error is at most a fixed fraction (δ) of their squared Frobenius norm. To this end, we exploit the fact that matrices whose row-spaces are uniformly distributed on $\mathrm{Gr}(m, n)$ may be modeled as matrices having iid zero-mean Gaussian elements. With this, we establish the following theorem

Theorem 2.5.1. For $2 \leq m \leq n$, let $\mathbf{A} \in \mathbb{R}^{m \times n}$ have iid $\mathcal{N}(0,1)$ entries. Then for $\delta \in [0, 0.125)$, and n is sufficiently large, there exists a positive constant $c(\delta)$ such that

$$\Pr(\mathcal{E}_{\mathcal{P}C_{m,n}}(\mathbf{A}) \leq \delta \|\mathbf{A}\|_F^2) = \mathcal{O}(e^{-c(\delta)\cdot mn}).$$

Proof. The proof is outlined in appendix section 2.11.2. □

Simply put, the content of Theorem 2.5.1 is that the proportion of large matrices (with row spaces uniformly distributed in $\mathrm{Gr}(m, n)$) that can be approximated to high accuracy (here, with small Frobenius approximation error) by partial circulant matrices is exponentially small in mn, the product of the matrix dimensions. In the next section we describe a more general framework designed to facilitate accurate partial circulant approximations in a number of practical applications.

2.6 Approximation with partial circulant matrices with post-processing

We now consider a more general approximation framework by approximating $\mathbf{A}$ as $\mathbf{A} \approx \mathbf{PSC}$ where $\mathbf{C} \in \mathcal{C}_n$, $\mathbf{S}$ is an $m' \times n$ matrix comprising a permuted subset of rows of an identity matrix, and $\mathbf{P} \in \mathbb{R}^{m \times m'}$ is a "post-processing" matrix. In this setting, the minimum approximation error is

$$\mathcal{E}_{\mathcal{P}C_{m,m',n}}(\mathbf{A}) = \min_{\mathbf{Z} \in \mathcal{P}C_{m,m',n}} \|\mathbf{A} - \mathbf{Z}\|_F^2. \tag{2.8}$$

Again evaluating $\mathcal{E}_{\mathcal{P}C_{m,m;,n}}(\mathbf{A})$ is a non-convex problem. However, quantifying the improvements offered by this expanded approximation model is possible when we make use of the following result.

Lemma 2.6.1. *For $\mathbf{A} \in \mathbb{R}^{m \times n}$, we have*

$$\mathcal{E}_{\mathcal{P}C_{m,m',n}}(\mathbf{A}) = \|\mathbf{A}\|_F^2 - \left[\max_{\mathbf{Z} \in \widetilde{\mathcal{P}C}_{m,m',n}} Tr\left(\mathbf{A}^T \mathbf{Z}\right) \right]^2,$$

where $\widetilde{\mathcal{P}C}_{m,m',n} = \left\{ \frac{\mathbf{Z}}{\|\mathbf{Z}\|_F} \middle| \mathbf{Z} \neq \mathbf{0}, \mathbf{Z} \in \mathcal{P}C_{m,m',n} \right\}.$

Proof. The proof is outline in appendix section 2.11.3 □

The quantity $\max_{\tilde{\mathbf{Z}} \in \widetilde{\mathcal{P}C}_{m,m',n}} \operatorname{Tr}\left(\mathbf{A}^T \tilde{\mathbf{Z}}\right)$ in the above result is analogous to the Rubik's score in the previous section. For a given $\mathbf{A}$, it is the key quantity that determines the quality of approximation provided by the circulant matrices with post-processing. In fact, it can be shown that the Rubik's score of a partial circulant matrix is exactly equal to

$$R(\mathbf{A}) = \max_{\mathbf{Z} \in \widetilde{\mathcal{P}C}_{m,n}} \operatorname{Tr}\left(\mathbf{A}^T \mathbf{Z}\right), \tag{2.9}$$

where $\widetilde{\mathcal{P}C}_{m,n} = \left\{ \frac{\mathbf{Z}}{\|\mathbf{Z}\|_F} \middle| \mathbf{Z} \neq \mathbf{0}, \mathbf{Z} \in \mathcal{P}C_{m,n} \right\}$. Also, since $\widetilde{\mathcal{P}C}_{m,n} \subset \widetilde{\mathcal{P}C}_{m,m',n}$ it implies that the Rubik's score satisfies

$$R(\mathbf{A}) \leq \max_{\tilde{\mathbf{Z}} \in \widetilde{\mathcal{P}C}_{m,m',n}} \operatorname{Tr}\left(\mathbf{A}^T \tilde{\mathbf{Z}}\right) \tag{2.10}$$

This implies that the approximation with partial circulant matrices that employ post-processing are guaranteed to be no worse than those with post-processing. This is not surprising given that partial circulant matrices with post-processing contain the partial circulant matrices as a special case. However, the exact improvement is difficult to quantify.

Our first attempt at quantifying this improvement involves the geometric insight that can be obtained upon row-wise vectorization of matrices involved in the problem. Upon row-wise vectorization the partial circulant matrices with post-processing matrix satisfy the following form

$$\operatorname{vec}(\mathbf{PSC})^T = \mathbf{c}^T \left[\sum_{j=1}^{m'} P_{1,j}\mathbf{R}^{f_j}, \cdots, \sum_{j=1}^{m'} P_{m,j}\mathbf{R}^{f_j}\right]$$

where $[f_1, \cdots, f_{m'}] \in \{0, \cdots, n-1\}^{m'}$ such that $f_i \neq f_j$ for $i \neq j$. For a fixed $\mathbf{P}$ the $\operatorname{vec}(\mathbf{PSC})^T$ lies along the row space of $\left[\sum_{j=1}^{m'} P_{1,j}\mathbf{R}^{f_j}, \cdots, \sum_{j=1}^{m'} P_{m,j}\mathbf{R}^{f_j}\right]$ and by varying $f_i's$ we can show that there are $\binom{n}{m'}m'!$ subspaces. Further, the lower bound $\left(\frac{n}{e}\right)^{m'} \leq \binom{n}{m'}m'!$ implies that the number of subspaces increase exponentially as we increase m'. Additionally, the rotations of the union of subspaces can be provided by varying $\mathbf{P}$. With these additional subspaces approximation with "post-processing" matrix is better as it provides denser cover for $\mathbb{R}^{mn}$. With these additional subspaces approximation with "post-processing" matrix is better. Figure 2.3 shows a representative geometric implication of approximation with "post-processing" matrix.

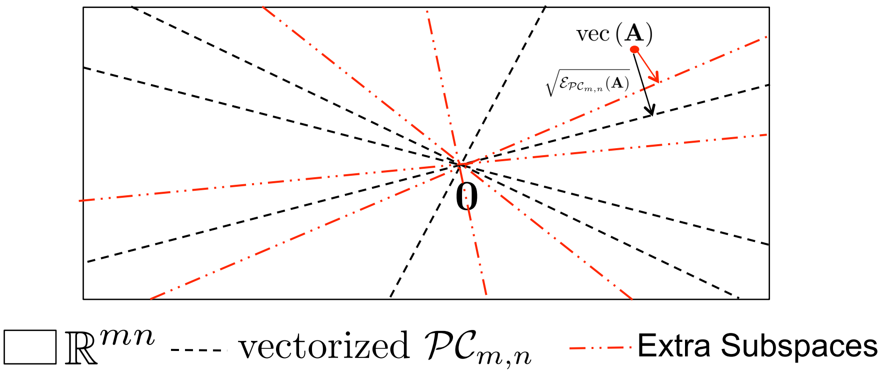

vec (A)
$\sqrt{\mathcal{E}_{\mathcal{PC}_{m,n}}(\mathbf{A})}$
0
$\mathbb{R}^{mn}$
vectorized $\mathcal{PC}_{m,n}$
Extra Subspaces

Figure 2.3: A representative figure showing the geometric implication of "post-processing" matrix in terms of denser union of subspaces providing better approximation.

Even though the above geometric interpretation provides nice insights into how the error decreases, it falls short of quantifying the improvement. Our second attempt towards exact quantification is based on a probabilistic approach. We assume that the matrix $\mathbf{A}$ is a Gaussian random matrix with i.i.d. entries following standard Gaussian distribution. The expected error incurred with random $\mathbf{A}$ generated in such a manner is bounded as stated in the following theorem.

Theorem 2.6.1. *Assuming that $\mathbf{A} \in \mathbb{R}^{m \times n}$ have iid $\mathcal{N}(0,1)$ entries the expected error is bounded as*

$$\mathbb{E}_{\mathbf{A}}\left(\mathcal{E}_{\mathcal{P}C_{m,m',n}}(\mathbf{A})\right) \leq mn - \left[\omega(\widetilde{\mathcal{P}C}_{m,m',n})\right]^2, \tag{2.11}$$

where $\omega(\widetilde{\mathcal{P}C}_{m,m',n})$ is the Gaussian width of the set $\widetilde{\mathcal{P}C}_{m,m',n}$ defined as

$$\omega(\widetilde{\mathcal{P}C}_{m,m',n}) = \mathbb{E}_{\mathbf{A}}\left(\max_{\mathbf{Z} \in \widetilde{\mathcal{P}C}_{m,m',n}} \operatorname{tr}\left(\mathbf{A}^{\mathrm{T}}\mathbf{Z}\right)\right). \tag{2.12}$$

Proof. The proof is outlined in section 2.11.4 □

Theorem 2.6.1 provides a fundamental insight that the average approximation error with partial circulant matrices with post-processing matrix is related to the Gaussian width of $\widetilde{\mathcal{P}C}_{m,m',n}$, and it decreases with increasing Gaussian width. The following lemma provides a lower bound on the Gaussian width of $\widetilde{\mathcal{P}C}_{m,m',n}$ as a function of m'.

Lemma 2.6.2. *The Gaussian width of the* $\widetilde{\mathcal{P}C}_{m,m',n}(\mathbf{A})$ *is bounded as*

$$\omega(\widetilde{\mathcal{P}C}_{m,m',n}) \geq \frac{mm'}{\sqrt{1+mm'}}. \tag{2.13}$$

Proof. The proof is outlined in section 2.11.5 □

A direct implication of Lemma 2.6.2 is that the expected error $\mathcal{P}C_{m,m',n}(\mathbf{A})$ is upper bounded as

$$\mathbb{E}\left(\mathcal{E}_{\mathcal{P}C_{m,m',n}}(\mathbf{A})\right) \leq mn\left(1 - \frac{m'}{n}\frac{mm'}{1+mm'}\right) \tag{2.14}$$

The above bound on expected error implies that as we increase m' the expected error decreases. Since for modest values of m, m' the term $\frac{mm'}{1+mm'} \approx 1$, we see that with each extra internal measurement expected the average approximation error decreases by m. In the next section we consider a more general framework designed to facilitate accurate partial circulant approximations in a number of practical applications.

2.7 The data driven approach

The underlying theme of the approximation techniques considered so far is that they try to approximate the given LDR matrix globally. However, in many practical applications where LDR methods are employed (e.g. PCA, Compressive Sensing etc.) the data to be processed are not arbitrary, but lie in some *restricted input domain* $\mathcal{X}$ (e.g., in a low-dimensional subspace, a union of low-dimensional subspaces, distinct clusters, etc.). In these cases, our approximation task simplifies to mimicking the action of $\mathbf{A}$ on these restricted inputs. This motivates us to propose the *data-driven* approach in which we are given the data matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$ whose columns are "representative" samples from the restricted input domain $\mathcal{X}$ for the problem of interest and we want to approximate the action of $\mathbf{A}$ only on $\mathbf{X}$.

Next, we discuss data-driven approaches first for the partial circulant matrices followed by the partial circulant matrices with a post-processing matrix.

2.7.1 Data driven approach for partial circulant matrices

The data-driven approach for partial circulant matrices involves approximating the action of matrix $\mathbf{A}$ on the given matrix of representative data $\mathbf{X} \in \mathbb{R}^{n \times p}$ using a matrix from the set $\mathcal{P}C_{m,n}$. For this purpose the "data-driven" approach here involves solving the following optimization problem

$$\min_{\mathcal{S}\in\mathcal{S}_m,\mathbf{C}\in\mathcal{C}_n} \|\mathbf{AX}-\mathbf{SCX}\|_F^2 \tag{2.15}$$

The above problem is a non-convex problem. We discuss an alternating minimization type algorithm for approximately solving this problem in Section 2.8.1.

2.7.2 Data driven approach for partial circulant matrices with post-processing

The presence of post-processing matrix allows for better approximation. The potential efficacy of the data-driven approach for partial circulant matrices with post-processing matrix can be made concrete by the following positive existence result.

Theorem 2.7.1. *Let $\mathbf{A} \in \mathbb{R}^{m\times n}$ be any fixed matrix, and let $\mathcal{X}$ be any finite set of n-dimensional unit-norm vectors. For any $\epsilon \in (0,1)$, there exists a post-processing $\mathbf{P} \in \mathbb{R}^{m\times m'}$, sampling matrix $\mathbf{S} \in \mathbb{R}^{m'\times n}$ comprised of rows of identity matrix $\mathbf{I}_n$, and circulant $\mathbf{C} \in \mathbb{C}^{n\times n}$ for which*

$$\sup_{\mathbf{x}\in\mathcal{X}} \|\mathbf{Ax}-\mathbf{PSCx}\|_2 \leq \epsilon\|\mathbf{A}\|_F,$$

provided that $m' > c_1\epsilon^{-2}\log(c_2 m|\mathcal{X}|)\log^4(n)$ where c_1 and c_2 are universal positive constants.

Proof. The proof is outlined in section 2.11.6 □

We note that the above result is for finite sized set but extensions to general sets can be derived using the covering number arguments. Motivated by this positive existence results we propose the "data-driven" approach which involves solving the following problem

$$\min_{\mathbf{P}\in\mathbb{R}^{m\times m'},\mathcal{S}\in\mathcal{S}_{m'},\mathbf{C}\in\mathcal{C}_n} \max_{i\in\{1,\cdots,n\}} \|\mathbf{Ax}_i-\mathbf{PSCx}_i\|_2^2, \tag{2.16}$$

where $\mathbf{X} \in \mathbb{R}^{n\times p}$ the given data matrix and $\mathbf{x}_i$ is the i^{th} training data point or the i^{th} column of the data matrix $\mathbf{X}$. For computational tractability we follow the bounded minimization approach. We use the following trivial upper bound on the maximum error over the training data

$$\max_{i\in\{1,\cdots,n\}} \|\mathbf{A}\mathbf{x}_i - \mathbf{PSC}\mathbf{x}_i\|_2^2 \leq \sum_{i=1}^{n} \|\mathbf{A}\mathbf{x}_i - \mathbf{PSC}\mathbf{x}_i\|_2^2 \tag{2.17}$$

and propose solving the following problem instead

$$\min_{\mathbf{P}\in\mathbb{R}^{m\times m'},\mathcal{S}\in\mathcal{S}_{m'},\mathbf{C}\in\mathcal{C}_n} \|\mathbf{AX} - \mathbf{PSCX}\|_F^2. \tag{2.18}$$

The above formulation minimizes essentially boils down to minimizing the error in average sense over the training data. In the above problem $\mathcal{S}_{m'}$ is discrete in nature. One needs to potentially solve the above problem for various values of m' and choose the one which gives the desired performance. We notice that for a fixed value of m' the resulting matrix $\mathbf{PS}$ obtained by this approach has exactly m' non-zero columns. Based on this observation, we propose a general framework in which we effectively combine the actions of the sampling and post processing matrices; we let $\mathbf{M} \triangleq \mathbf{PS}$, and seek *column sparsity* in $\mathbf{M}$ using the $\|\mathbf{M}\|_{2,1}$ regularization term as follows

$$\min_{\mathbf{M}\in\mathbb{R}^{m\times n},\mathbf{C}\in\mathcal{C}_n} \|\mathbf{AX} - \mathbf{MCX}\|_F^2 + \lambda\|\mathbf{M}\|_{2,1} + \mu\|\mathbf{C}\|_F^2, \tag{2.19}$$

where $\lambda > 0, \mu > 0$ are the regularization parameters. The regularization term $\|\mathbf{C}\|_F^2$ is needed to fix scaling ambiguities introduced due to the matrix product term $\mathbf{MC}$. In the above formulation the non-zero columns m' vary with λ. This problem is also non-convex. An alternating algorithm minimization type algorithm to approximately solve this problem is discussed in Section 2.8.2.

2.8 Algorithms for the data driven approach

In this section, we discuss the algorithm for the problems arising in the data-driven approach of matrix approximation. Since both the problems are non-convex we present alternating minimization based algorithms for both the cases.

2.8.1 Data driven approach for partial circulant matrices

As discussed earlier in Section 2.7.1, the data-driven approach for approximating the action of given LDR matrix $\mathbf{A}$ on the given matrix of representative data $\mathbf{X} \in \mathbb{R}^{n\times p}$

using a partial circulant matrix involves solving the problem in (2.15). This problem is jointly non-convex in $\mathbf{C}$ and $\mathbf{S}$ due the discrete nature of the set $\mathcal{S}_m$ and the matrix multiplication term $\mathbf{SC}$. So we propose an alternating minimization type approach to approximately solve it. Starting with initial feasible points $\mathbf{S}^{(0)}$ and $\mathbf{C}^{(0)}$ we alternatively solve the following problem till convergence as shown in Algorithm 1.

$$\mathbf{C}^{(t)} = \arg\min_{\mathbf{C}\in\mathcal{C}_n} \ \|\mathbf{AX} - \mathbf{S}^{(t-1)}\mathbf{CX}\|_F^2, \tag{2.20}$$

$$\mathbf{S}^{(t)} = \arg\min_{\mathbf{S}\in\mathcal{S}_m} \ \|\mathbf{AX} - \mathbf{SC}^{(t)}\mathbf{X}\|_F^2. \tag{2.21}$$

Next we discuss the above two update steps.

C update step

The $\mathbf{C}$ update step in 2.20 can be converted to a least squares problem in the first row $\mathbf{c}$ of the circulant matrix. Since the matrix $\mathbf{S}^{(t-1)}$ selects some m-unique rows of $\mathbf{C}$ using the representation of $\mathbf{C}$ in (2.1) the objective cost in 2.20 can be written as

$$\begin{aligned}\|\mathbf{AX}-&\mathbf{S}^{(t-1)}\mathbf{CX}\|_F^2\\ &= \sum_{i=1}^{m} \|\mathbf{A}_{i,:}\mathbf{X} - \mathbf{c}^T\mathbf{R}^{f_i^{(t-1)}}\mathbf{X}\|_2^2,\\ &= \|\mathbf{AX}\|_F^2 + m\mathbf{c}^T\mathbf{XX}^T\mathbf{c} - 2\sum_{i=1}^{m}\mathbf{A}_{i,:}\mathbf{XR}^{f_i^{(t-1)}}\mathbf{X}^T\mathbf{c}.\end{aligned}$$

where $f_i^{(t-1)} \in \{0, \cdots, n-1\}$ whose value depends on the i^{th} row $\mathbf{S}^{(t-1)}$ such that if i^{th} row selects k^{th} row of $\mathbf{C}$ then $f_i^{(t-1)} = k-1$. Taking the derivative with respect to $\mathbf{c}$ and equating it to zero the optimal $\mathbf{c}$ is given by

$$\mathbf{c}^{(t)} = \left(\mathbf{XX}^T\right)^{\dagger}\left(\mathbf{X}\left(\sum_{i=1}^{m}\frac{1}{m}\mathbf{L}^{f_i^{(t-1)}}\mathbf{X}^T\mathbf{A}_{i,:}^T\right)\right). \tag{2.22}$$

The full circulant matrix can be from $\mathbf{c}^{(t)}$ above using (2.1). We note that $\left(\mathbf{XX}^T\right)^{\dagger}$ and $\mathbf{A}_{i,:}\mathbf{X}$ do not change with different iterations so they can be pre-computed once and stored at the start of the alternating minimization algorithm.

S update step

Even though the $\mathbf{S}$ update step appears to be combinatorial non-convex problem it can be converted into a linear program and can be solved efficiently. The exact equivalent linear program is described in the following theorem:

Theorem 2.8.1. *The* $\mathbf{S}^{(t)}$ *in* (2.21) *is equal to the first* m*-rows of* $\mathbf{P}^*$ *where*

$$\begin{array}{rc} \mathbf{P}^* = \arg\min\limits_{\mathbf{P}\in\mathbb{R}^{n\times n}} & \sum_{i=1}^{n}\sum_{i=1}^{n} W_{ij}^{(t)} P_{ij} \\ & \sum_{i=1}^{n} P_{ij} = 1, \forall j \\ & \sum_{j=1}^{n} P_{ij} = 1, \forall i \\ & P_{ij} \geq 0, \forall i, j \end{array}$$

where

$$W_{ij}^{(t)} = \begin{cases} \|\mathbf{A}_{i,:}\mathbf{X} - \mathbf{C}_{j,:}^{(t)}\mathbf{X}\|_2^2 & 1 \leq i \leq m, \forall j \\ 0 & i > m, \forall j. \end{cases}$$

Proof. The proof is outlined in section 2.11.9. □

We note that for the equivalent linear program in the above theorem $\mathbf{A}_{i,:}\mathbf{X}$ can be precomputed and stored at the beginning of alternating minimization. However, here the value of $\mathbf{C}^{(t)}\mathbf{X}$ changes with every iteration. Since the matrix $\mathbf{C}^{(t)}$ is circulant, one may utilize its FFT based fast implementation for calculating it. We note that entire alternating algorithm can be run without actually constructing full $\mathbf{C}$ because FFT based circulant matrix multiplication can be carried out just by the knowledge of first row of $\mathbf{C}$.

Based on the above discussed updates for the two sub-problems the final alternating minimization for data-driven approximation of a partial circulant matrix is given in Algorithm 1.

Algorithm 1 "Data-Driven" Partial Circulant Approximation.

```
Inputs: LDR matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$, parameter$\epsilon > 0$,
        Matrix of "representative" data $\mathbf{X} \in \mathbb{R}^{n \times p}$,
Precompute: $\left(\mathbf{X}\mathbf{X}^T\right)^\dagger, \mathbf{A}\mathbf{X}, \mathrm{fft}(\mathbf{X})$.
Initialize: $\mathbf{S}^{(0)}$ = first $m$-rows of $n \times n$ identity matrix.
        $\mathrm{obj}^{(0)} = \|\mathbf{A}\mathbf{X}\|_F^2$,
  repeat
    Compute $f_i^{(t-1)} = k - 1$ where $k$ is the location of 1 in $i^{th}$ row $\mathbf{S}_{i,:}^{(t-1)}$.
    Update $\mathbf{c}^{(t)} = \left(\mathbf{X}\mathbf{X}^T\right)^\dagger \left(\mathbf{X}\left(\sum_{i=1}^m \frac{1}{m}\mathbf{L}^{f_i^{(t-1)}}\mathbf{X}^T\mathbf{A}_{i,:}^T\right)\right)$
    Compute $\mathbf{C}^{(t)}\mathbf{X} = \mathrm{fft}\left(\mathrm{Diag}\left(\mathrm{ifft}\left(\mathbf{c}^{(t)}\right)\right)\mathrm{fft}\left(\mathbf{X}\right)\right)$ and
        $W_{ij}^{(t)} = \begin{cases} \|\mathbf{A}_{i,:}\mathbf{X} - \mathbf{C}_{j,:}^{(t)}\mathbf{X}\|_2^2 & 1 \le i \le m, \forall j \\ 0 & i > m, \forall j. \end{cases}$
    $\mathbf{S}^{(t)}$ = first $m$-rows of $\mathbf{P}^{(t)}$ where
        $\mathbf{P}^{(t)} = \arg\min_{\mathbf{P} \in \mathbb{R}^{n \times n}} \sum_{i=1}^n \sum_{i=1}^n W_{ij}^{(t)} P_{ij}$
            $\sum_{i=1}^n P_{ij} = 1, \forall j$
            $\sum_{j=1}^n P_{ij} = 1, \forall i$
            $P_{ij} \ge 0, \forall i, j$
    $\mathrm{obj}^{(t)} = \|\mathbf{A}\mathbf{X} - \mathbf{S}^{(t)}\mathbf{C}^{(t)}\mathbf{X}\|_F^2$
  until $\mathrm{obj}^{(t)} - \mathrm{obj}^{(t-1)} \le \epsilon \cdot \mathrm{obj}^{(t-1)}$
Compute $\mathbf{C}^{(t)}$ using $\mathbf{c}^{(t)}$ using (2.1)
Output: $\mathbf{S}^* = \mathbf{S}^{(t)}, \mathbf{C}^* = \mathbf{C}^{(t)}$
```

2.8.2 Data driven approach for partial circulant matrices with post-processing

As discussed earlier in Section 2.7.2, the data-driven approach for approximating the action of given LDR matrix $\mathbf{A}$ on the given matrix of representative data $\mathbf{X} \in \mathbb{R}^{n \times p}$ using partial circulant matrix with post-processing matrix involves solving problem in (2.19).This problem is jointly non-convex in $\mathbf{M}$ and $\mathbf{C}$ due to the matrix multiplication term. However, for a fixed value of $\mathbf{M}$ the problem is convex in $\mathbf{C}$ and vice versa. So we propose an alternating minimization based approach to solve it. Starting with initial feasible $\mathbf{M}^{(0)}, \mathbf{C}^{(0)}$ the following two sub-problems are solved alternatively until convergence

$$\mathbf{C}^{(t)} = \arg \min_{\mathbf{C} \in \mathcal{C}_n} \|\mathbf{AX} - \mathbf{M}^{(t-1)}\mathbf{CX}\|_F^2 + \mu \|\mathbf{C}\|_F^2 \tag{2.23}$$

$$\mathbf{M}^{(t)} = \arg \min_{\mathbf{M} \in \mathbb{R}^{m \times n}} \|\mathbf{AX} - \mathbf{MC}^{(t)}\mathbf{X}\|_F^2 + \lambda \|\mathbf{M}\|_{2,1} \tag{2.24}$$

Next we discuss the above two update steps.

C update

The $\mathbf{C}$ update step is a convex problem and it can be converted to a least squares problem in the first row of $\mathbf{C}$ similar to as was done for the case without post-processing matrix. However, there is a critical difference here that the Hessian depends on the $\mathbf{M}^{(t-1)}$ so closed form solution may be prohibitive to obtain for large n because it involves calculating a pseudo inverse of a $n \times n$ matrix. So we propose a projected first order gradient type algorithm for $\mathbf{C}$ update which involves taking a gradient step followed by projection onto set of circulant matrices. Projection of a given matrix $\mathbf{U} \in \mathbb{R}^{n \times n}$ onto the set of circulant matrices can be posed as the following problem

$$\mathbf{C}^{\mathbf{U}} = \arg \min_{\mathbf{C} \in \mathbf{C}_n} \|\mathbf{U} - \mathbf{C}\|_F^2 \tag{2.25}$$

Using the structure of the circulant matrices it is easy to see that the first row of the projected circulant matrix is given by

$$\mathbf{c}^{\mathbf{U}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{R}^{i-1} \mathbf{U}_{i,:}^T. \tag{2.26}$$

Using the above first row the full circulant matrix $\mathbf{C}^{\mathbf{U}}$ can be obtained by using (2.1). Based on the above projection step we propose accelerated projected gradient algorithm for $\mathbf{C}$ update. This algorithm is based on the accelerated proximal method proposed in [59]. The overall algorithm is shown Algorithm 2[2] [3] .

[2] Here 2 η is the gradient Lipschitz constant of the objective function in the optimization problem. The Hessian of objective function with respect to $\mathbf{C}$ is $2\mu\mathbf{I}_n + 2(\mathbf{XX}^T) \otimes (\mathbf{MM}^T)$ so the gradient Lipschitz constant is given by $2\mu + 2\|\mathbf{X}\|_2^2 \|\mathbf{M}\|_2^2$

[3] This Algorithm can also be used for the case without post-processing matrix as well by calling it with $\mu = 0$ and $\mathbf{M} = \mathbf{S}$.

Algorithm 2 Accelerated projected gradient descent $\mathbf{C}$ update step: $\min_{\mathbf{C}\in\mathcal{C}_n} \|\mathbf{AX} - \mathbf{MCX}\|_F^2 + \mu\|\mathbf{C}\|_F^2$.

```
Inputs: $\mathbf{A} \in \mathbb{R}^{m\times n}$, $\mathbf{M} \in \mathbb{R}^{m\times n}$, $\mathbf{X} \in \mathbb{R}^{n\times p}$, initial $\mathbf{C}^{(0)}$, $\mu > 0, \epsilon > 0$
Initialize: $\mathbf{Z}^{(1)} = \mathbf{C}^{(0)}$, $\eta = 2\mu + 2\|\mathbf{M}\|_2^2\|\mathbf{X}\|_2^2$
Repeat: for $t \geq 1$
  Compute gradient:
    $\mathbf{G}^{(t)} = 2\mathbf{M}^T\left(\mathbf{MZ}^{(t)}\mathbf{X} - \mathbf{AX}\right)\mathbf{X}^T + 2\mu\mathbf{Z}^{(t)}$
  Gradient step:
    $\mathbf{U} = \mathbf{Z}^{(t)} - \eta\mathbf{G}^{(t)}$
  Projection on $\mathcal{C}_n$:
   Find the first row: $\mathbf{c}^{(t)} = \frac{1}{n}\sum_{i=1}^n \mathbf{R}^{i-1}\mathbf{U}_{i,:}^T$
   Construct $\mathbf{C}^{(t)}$ from $\mathbf{c}^{(t)}$ using (2.1)
  Accelerate: $\mathbf{Z}^{(t+1)} = \mathbf{C}^{(t)} + \frac{t}{t+3}\left(\mathbf{C}^{(t)} - \mathbf{C}^{(t-1)}\right)$
Until: $\|\mathbf{C}^{(t)} - \mathbf{C}^{(t-1)}\|_F \leq \epsilon \cdot \|\mathbf{C}^{(t-1)}\|_F$
Output: $\mathbf{C}^* = \mathbf{C}^{(t)}$
```

M update

The **M** update step is a standard group Lasso problem that can be solved using existing software (e.g., SLEP [60]).

Using the **C** and **M** update steps the final alternating minimization algorithm is shown in Algorithm 3. The initialization is motivated by the initialization used for alternating minimization algorithm presented in [61]. We also note that one may use $\mathbf{C}^{(t-1)}$ as initialization while calling Algorithm 2 for **C** update step. This significantly speeds up the convergence.

Algorithm 3 "Data-Driven" Partial Circulant Approximation with post-processing matrix.

```
Inputs: LDR matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$, parameters $\lambda, \mu, \epsilon > 0$,
        Matrix of "representative" data $\mathbf{X} \in \mathbb{R}^{n \times p}$,
Initialize: $\mathbf{M}^{(0)} = \mathbf{U}\Sigma$ (from the SVD $\mathbf{AX} = \mathbf{U}\Sigma\mathbf{V}^T$)
        $\text{obj}^{(0)} = \|\mathbf{AX}\|_F^2$
  repeat
    $\mathbf{C}^{(t)} = \arg\min_{\mathbf{C} \in \mathcal{C}_n} \|\mathbf{AX} - \mathbf{M}^{(t-1)}\mathbf{CX}\|_F^2 + \mu\|\mathbf{C}\|_F^2$
    $\mathbf{M}^{(t)} = \arg\min_{\mathbf{M} \in \mathbb{R}^{m \times n}} \|\mathbf{AX} - \mathbf{MC}^{(t)}\mathbf{X}\|_F^2 + \lambda\|\mathbf{M}\|_{2,1}$
    $\text{obj}^{(t)} = \|\mathbf{AX} - \mathbf{M}^{(t)}\mathbf{C}^{(t)}\mathbf{X}\|_F^2 + \mu\|\mathbf{C}^{(t)}\|_F^2 + \lambda\|\mathbf{M}^{(t)}\|_{2,1}$
  until $\text{obj}^{(t)} - \text{obj}^{(t-1)} \leq \epsilon \cdot \text{obj}^{(t-1)}$
Output: $\mathbf{M}^* = \mathbf{M}^{(t-1)}, \mathbf{C}^* = \mathbf{C}^{(t-1)}$
```

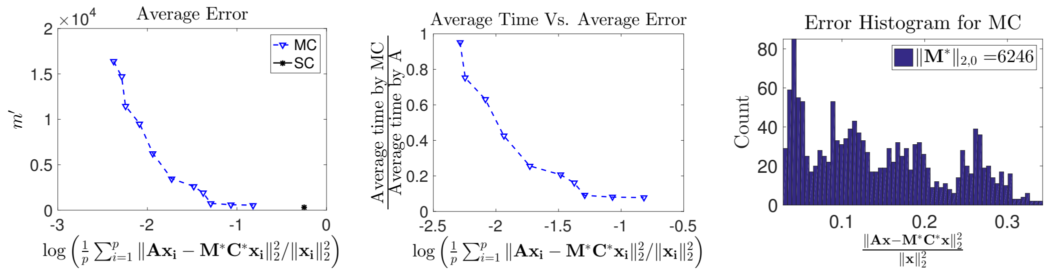


Figure 2.4: Data-driven approximation of matrix comprising of top 300 principal components of a training set of images from COIL-20 database. The first panel (left to right) contains log average relative approximation error vs. m' with various circulant approximations to the given matrix. The second panel plots shows the ratio of average time taken by these circulant approximation and $\mathbf{A}$ vs. the log average relative approximation error. The third panel contains the histograms of errors by the circulant approximation matrix obtained by Algorithm 3.

2.9 Experimental evaluation

We evaluate these approaches using the processed COIL-20 image database available at `http://www.cs.columbia.edu/CAVE/software/softlib/coil-20.php`. This database contains 128 × 128 images which were vectorized to obtain a data matrix $\mathbf{X}$ whose columns represent 1440 vectorized images from the dataset. We took rows of $\mathbf{A}$ as the top 300 principal component vectors of the training data. In Algorithm 3, we use $\mu = 0.1$ and vary λ to obtain $\mathbf{M}^*$ and $\mathbf{C}^*$ for each value of λ, and quantified the normalized error on the training set vs. column sparsity of $\mathbf{M}^*$. We also plot the results for circulant approximation to the matrix without the post-processing obtained using Algorithm 1.

The first panel (left to right) in Figure 2.4 plots m' vs. log average relative error for "data-driven" approximations with and without post-processing matrix. We can see from the plot that the approximation with post-processing matrix (shown in blue colored dotted line with triangle shaped marker) incurs far less error as compared to approximation without post-processing matrix (shown in black star shaped marker). This plot demonstrates the superiority of approximations with post-processing matrix relative to that without post-processing matrix.

The second panel (left to right) in Figure 2.4 give insight into the relative time taken by matrix approximations shown as blue markers in the plot in first panel as compared the original LDR matrix $\mathbf{A}$. It plots the ratio of average time taken by $\mathbf{A}$ and by its "data-driven" approximation with post-processing matrix obtained from Algorithm 3 versus the log average relative error. The average time was obtained by averaging over the training data set.[4] We can see that the multiplication with approximations obtained by Algorithm 3 is faster than the given LDR matrix $\mathbf{A}$. The speed is due to the FFT based implementation of matrix vector multiplication. We can see that as m' increases the speed of matrix vector multiplication decreases but the log average relative error also decreases. This plot illustrates the inherent speed vs. accuracy trade-off.

For a representative λ corresponding to the sparsity of $m' = 6246$, we compute the histogram of the normalized approximation errors ($\|\mathbf{Ax} - \mathbf{MCx}\|_2^2 / \|\mathbf{Ax}\|_2^2$) for each point in the training data set. The histogram is plotted in third panel of the Figure 2.4. Most of relative errors are small which demonstrates that our approach provides fairly accurate approximation.

[4] The matrix vector multiplication was conducted by C programming language based implementation that uses FFT algorithm. The average time vs. average error plots were obtained by averaging over 10 trials.

2.10 Summary

We investigated the problem of approximating an arbitrary LDR matrix via various partial circulant structured matrices, presented several fundamental results, and evaluated numerically a data-driven partial circulant approximation approach. Future directions of research include extension of the basic analytical framework developed here to other structured matrix approximations with low implementation complexities (e.g., sparse matrices and fast Johnson-Lindenstrauss embeddings [62, 63]); this is a topic of our ongoing work. See section 7.1 for more discussion on this and several other possible future directions of research.

2.11 Appendix

2.11.1 Proof of lemma 2.4.1

Proof. We first write $\mathcal{E}_{\mathcal{P}C_{m,n}}(\mathbf{A})$ in (2.5) equivalently as

$$\mathcal{E}_{\mathcal{P}C_{m,n}}(\mathbf{A}) = \min_{\mathbf{S}\in\mathcal{S}_m} \min_{\mathbf{C}\in\mathcal{C}_n} \;\|\mathbf{A} - \mathbf{S}\mathbf{C}\|_F^2. \tag{2.27}$$

Each choice of $\mathbf{S} \in \mathcal{S}_m$ corresponds to a integer valued vector $\mathbf{f}$ from the set $\mathcal{F}$ defined in Lemma 2.4.1, because including the i-th row of $\mathbf{I}_n$ in $\mathbf{S}$ corresponds to selecting the i-th row of $\mathbf{C}$, which is given from (2.4) by $\mathbf{c}^T\mathbf{R}^{i-1}$. Using this we may parameterize the choice of $\mathbf{S}$ in terms of $\mathbf{f}$, and rewrite the objective function in (2.27) as

$$\|\mathbf{A} - \mathbf{S}\mathbf{C}\|_F^2 = \sum_{i=1}^{m} \|\mathbf{A}_{i,:}\|_2^2 + \mathbf{c}^T\mathbf{R}^{f_i}\mathbf{L}^{f_i}\mathbf{c} - 2\mathbf{A}_{i,:}\mathbf{L}^{f_i}\mathbf{c}.$$

Thus, since $\mathbf{R}^{f_i}\mathbf{L}^{f_i} = \mathbf{I}_n$, we have

$$\mathcal{E}_{\mathcal{P}C_{m,n}}(\mathbf{A}) = \min_{\mathbf{f}\in\mathcal{F},\mathbf{c}\in\mathbb{R}^n} \;\|\mathbf{A}\|_F^2 + m\|\mathbf{c}\|_2^2 - 2\left[\sum_{i=1}^{m} \mathbf{A}_{i,:}\mathbf{L}^{f_i}\right]\mathbf{c}. \tag{2.28}$$

We first minimize with respect to $\mathbf{c}$, keeping $\mathbf{f} \in \mathcal{F}$ fixed. This is an unconstrained strictly convex quadratic problem whose minimum can be obtained by equating the gradient (with respect to $\mathbf{c}$) to zero. Substituting the minimizer $\mathbf{c}^* = \frac{1}{m}\sum_{i=1}^{m}\mathbf{R}^{f_i}(\mathbf{A}_{i,:})^T$ into (2.28), and simplifying, we obtain

$$\mathcal{E}_{\mathcal{P}C_{m,n}}(\mathbf{A}) = \min_{\mathbf{f}\in\mathcal{F}} \|\mathbf{A}\|_F^2 - \frac{1}{m}\|\sum_{i=1}^{m}\mathbf{A}_{i,:}\mathbf{L}^{f_i}\|_2^2,$$

which gives (2.6), using the definition (2.7) of the *Rubik's Score.*

□

2.11.2 Proof of theorem 2.5.1

Proof. Note that $|\mathcal{F}| = n!/(n-m)!$ (since it is just the number of ways of choosing m elements out of n without replacement), and we have (trivially) that $n!/(n-m)! < n^m$, so

$$\begin{aligned} &\Pr(\mathcal{E}_{\mathcal{P}C_{m,n}}(\mathbf{A}) \leq \delta\|\mathbf{A}\|_F^2) \\ &\quad= \Pr\left((1-\delta)\|\mathbf{A}\|_F^2 \leq \sup_{f\in\mathcal{F}} \frac{\|\sum_{i=1}^m \mathbf{A}_{i,:}\mathbf{L}^{f_i}\|_2^2}{m}\right) \\ &\quad= \Pr\left(\bigcup_{f\in\mathcal{F}}\left[(1-\delta)\|\mathbf{A}\|_F^2 \leq \frac{\|\sum_{i=1}^m \mathbf{A}_{i,:}\mathbf{L}^{f_i}\|_2^2}{m}\right]\right) \\ &\quad\leq n^m \Pr\left((1-\delta)\|\mathbf{A}\|_F^2 \leq \frac{\|\sum_{i=1}^m \mathbf{A}_{i,:}\mathbf{L}^{f_i}\|_2^2}{m}\right), \end{aligned}$$

where the last step follows from union bounding.

Further, let $\mathbf{a} = [\mathbf{A}_{i,:}\ \mathbf{A}_{2,:}\ \ldots\ \mathbf{A}_{m,:}]^T \in \mathbb{R}^{mn}$ and $\tilde{\mathbf{R}}_{\mathbf{f}} = [\mathbf{R}^{f_1},\ \mathbf{R}^{f_2},\ \ldots, \mathbf{R}^{f_m}] \in \mathbb{R}^{n\times mn}$ with this we have $\|\mathbf{A}\|_F^2 = \mathbf{a}^T\mathbf{a}$ and

$$\frac{\left\|\sum_{i=1}^m \mathbf{A}_{i,:}\mathbf{L}^{f_i}\right\|_2^2}{m} = \frac{\|\tilde{\mathbf{R}}_{\mathbf{f}}\mathbf{a}\|_2^2}{m} = \mathbf{a}^T\left(\frac{\tilde{\mathbf{R}}_{\mathbf{f}}^T\tilde{\mathbf{R}}_{\mathbf{f}}}{m}\right)\mathbf{a}.$$

It is easy to check that

$$\left(\frac{\tilde{\mathbf{R}}_{\mathbf{f}}^T\tilde{\mathbf{R}}_{\mathbf{f}}}{m}\right)\tilde{\mathbf{R}}_{\mathbf{f}}^T = \tilde{\mathbf{R}}_{\mathbf{f}}^T,$$

which implies that the columns of $\tilde{\mathbf{R}}_{\mathbf{f}}^T$ are eigenvectors of the matrix corresponding to eigenvalue 1. Moreover, since $(\tilde{\mathbf{R}}_{\mathbf{f}}^T\tilde{\mathbf{R}}_f/m)$ is symmetric, it admits an eigendecomposition $(\tilde{\mathbf{R}}_{\mathbf{f}}^T\tilde{\mathbf{R}}_{\mathbf{f}}/m) = \mathbf{U_f}\Sigma_{\mathbf{f}}\mathbf{U_f}^T$ with $\mathbf{U_f}$ orthonormal and $\Sigma_{\mathbf{f}}$ diagonal, and since it is rank n, $\Sigma_{\mathbf{f}}$ has exactly n entries being 1 (and the rest 0). Incorporating this into the analysis above, we obtain that

$$\begin{aligned} &\Pr\left((1-\delta)\|\mathbf{A}\|_F^2 \leq \frac{\|\sum_{i=1}^m \mathbf{A}_{i,:}\mathbf{L}^{f_i}\|_2^2}{m}\right) \\ &\quad= \Pr\left(\mathbf{a}^T\mathbf{U_f}\ ((1-\delta)\mathbf{I}_{mn} - \Sigma_{\mathbf{f}}))\ \mathbf{U_f}^T\mathbf{a} \leq 0\right) \\ &\quad= \Pr\left(\tilde{\mathbf{a}}^T\ \left((1-\delta)\,\mathbf{I}_{mn} - \Sigma_{\mathbf{f}}\right)\ \tilde{\mathbf{a}} \leq 0\right), \end{aligned}$$

where the components of $\tilde{\mathbf{a}} = \mathbf{U}_{\mathbf{f}}\mathbf{a}$ are iid $\mathcal{N}(0,1)$ due to the unitary invariance of the Gaussian distribution. Thus, with a slight overloading of notation, we may write that

$$Pr(\mathcal{E}_{\mathcal{P}C_{m,n}}(\mathbf{A}) \leq \delta\|\mathbf{A}\|_F^2) \quad \leq \quad n^m \Pr\left(\sum_{i=2}^{m} \|\mathbf{A}_{i,:}\|_2^2 \leq \delta \sum_{i=1}^{m} \|\mathbf{A}_{i,:}\|_2^2\right). \tag{2.29}$$

At this point, we note that vectorizing (row-wise) a random matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$ having iid zero-mean Gaussian elements yields an mn-dimensional vector whose direction is selected uniformly at random from $\mathrm{Gr}(1, mn)$ (the space of 1-dimensional subspace in $\mathbb{R}^{mn}$). Thus, $\sum_{i=1}^{m} \|\mathbf{A}_{i,:}\|_2^2$ quantifies the length of the vector, while $\sum_{i=2}^{m} \|\mathbf{A}_{i,:}\|_2^2$ describes the energy retained after projecting the vector onto the fixed $n(m-1)$-dimensional subspace spanned by the last $n(m-1)$ coordinates. It follows that the probability on the right-hand side of (2.29) may be interpreted in terms of the energy retained after projecting a *fixed* mn-dimensional unit-normed vector onto a subspace selected uniformly at random from $\mathrm{Gr}(n, mn)$. To quantify this, we use the following.

Lemma 2.11.1 (From Thm. 2.14 of [64]). *Let* $\mathbf{v}$ *be a fixed unit-normed vector in* $\mathbb{R}^d$, W *a randomly oriented* k*-dim. subspace, and* $\mathbf{w}$ *the projection of* $\mathbf{v}$ *onto* W. *For* $\epsilon \in [0,1]$,

$$\Pr\left(\|\mathbf{w}\|_2 \leq (1-\varepsilon)\sqrt{k/d}\right) \leq 3e^{-k\varepsilon^2/64}.$$

Using this result (with $k = n(m-1)$ and $d = mn$) we obtain that for $\delta < 1 - 1/m$,

$$\begin{aligned}
&\Pr(\mathcal{E}_{\mathcal{P}C_{m,n}}(\mathbf{A}) \leq \delta\|\mathbf{A}\|_F^2) \\
&\leq \quad n^m \Pr\left(\frac{\sum_{i=2}^{m} \|\mathbf{A}_{i,:}\|_2^2}{\sum_{i=1}^{m} \|\mathbf{A}_{i,:}\|_2^2} \leq \delta\right) \\
&\leq \quad 3n^m e^{-\frac{n(m-1)}{64}\left(1-\sqrt{\frac{\delta}{1-\frac{1}{m}}}\right)^2} \\
&\leq \quad 3\ e^{m\log n - \frac{n(m-1)}{64}\left(1-2\sqrt{\frac{\delta}{1-\frac{1}{m}}}\right)}.
\end{aligned}$$

It is straightforward to verify that for $\delta < 1/8 = 0.125$ and n sufficiently large, so that $n/\log n > 128/(1-2\sqrt{2\delta})$, there exists positive $c(\delta) < \left(1-2\sqrt{2\delta}\right)/128 - (\log n)/n$, for which we have

$$m\log n - \frac{n(m-1)}{64}\left(1 - 2\sqrt{\frac{\delta}{1-\frac{1}{m}}}\right) \leq -c(\delta)mn.$$

In this case we have $\Pr(\mathcal{E}_{\mathcal{P}C_{m,n}}(\mathbf{A}) \leq \delta\|\mathbf{A}\|_F^2) \leq 3e^{-c(\delta)mn}$. □

2.11.3 Proof of lemma 2.6.1

Proof. The error $\mathcal{E}_{\mathcal{P}C_{m,m',n}}(\mathbf{A})$ can written as

$$\mathcal{E}_{\mathcal{P}C_{m,m',n}}(\mathbf{A}) = \|\mathbf{A}\|_F^2 + \min_{\mathbf{Z}\in\mathcal{P}C_{m,m',n}} \|\mathbf{Z}\|_F^2 - 2\mathrm{Tr}\left(\mathbf{A}^T\mathbf{Z}\right) \tag{2.30}$$

The set $\mathcal{P}C_{m,m',n}$ could be of arbitrary scaling, i.e., if $\mathbf{A} \in \mathcal{P}C_{m,m',n}$ them $\alpha\mathbf{A}$ is also in $\mathcal{P}C_{m,m',n}$ for all $\alpha \in \mathbb{R}$. The optimization over the set $\mathcal{P}C_{m,m',n}$ can be broken into scaling and direction. For this purpose we introduce the set of unit Frobenius norm matrices obtained from matrices in $\mathcal{P}C_{m,m',n}$ as

$$\widetilde{\mathcal{P}C}_{m,m',n} = \left\{\mathbf{Z}/\|\mathbf{Z}\|_F \middle| \mathbf{Z} \neq \mathbf{0}, \mathbf{Z} \in \mathcal{P}C_{m,m',n}\right\}.$$

In terms of $\widetilde{\mathcal{P}C}_{m,m',n}$ we can write $\mathcal{P}C_{m,m',n}$ as

$$\mathcal{P}C_{m,m',n} = \left\{\alpha\mathbf{Z} \middle| \alpha \geq 0, \mathbf{Z} \in \widetilde{\mathcal{P}C}_{m,m',n}\right\}.$$

Leveraging this fact we further simplify $\mathcal{E}_{\mathcal{P}C_{m,m',n}}(\mathbf{A})$ as follows

$$\begin{aligned}\mathcal{E}_{\mathcal{P}C_{m,m',n}}(\mathbf{A}) &= \|\mathbf{A}\|_F^2 + \min_{\mathbf{Z}\in\widetilde{\mathcal{P}C}_{m,m',n}} \min_{\alpha\geq 0} \alpha^2\|\mathbf{Z}\|_F^2 - 2\alpha\mathrm{Tr}\left(\mathbf{A}^T\mathbf{Z}\right) \\ &= \|\mathbf{A}\|_F^2 + \min_{\mathbf{Z}\in\widetilde{\mathcal{P}C}_{m,m',n}} \min_{\alpha\geq 0} \alpha^2 - 2\alpha\mathrm{Tr}\left(\mathbf{A}^T\mathbf{Z}\right), \end{aligned} \tag{2.31}$$

where last step utilizes the condition that $\|\mathbf{Z}\|_F = 1$ for all $\mathbf{Z} \in \widetilde{\mathcal{P}C}_{m,m',n}$. Next we solve the inner minimization with respect to α for a fixed $\mathbf{Z} \in \mathcal{P}C_{m,m',n}$. This problem is a convex optimization problem with strictly convex objective so the first order KKT conditions are necessary and sufficient for optimality [65]. For KKT conditions we first write the Lagrangian

$$\mathcal{L}(\alpha, \lambda) = \alpha^2 - 2\alpha\mathrm{Tr}\left(\mathbf{A}^T\mathbf{Z}\right) - \lambda\alpha, \tag{2.32}$$

where $\lambda \geq 0$ is Lagrangian variable for the condition $\alpha \geq 0$. The first KKT conditions are given by

1. $2\alpha^* - 2\mathrm{Tr}\left(\mathbf{A}^T\mathbf{Z}\right) - \lambda^* = 0,$

2. $\lambda^*\alpha^* = 0,$

3. $\lambda^* \geq 0, \alpha^* \geq 0$,

where α^*, λ^* are their optimal values. From the first KKT condition we have

$$\alpha^* = \mathrm{Tr}\left(\mathbf{A}^T\mathbf{Z}\right) + \lambda^*/2, \tag{2.33}$$

and the second KKT condition implies that if $\lambda^* > 0$ then $\alpha^* = 0$ which further implies that $\lambda^* = -2\mathrm{Tr}\left(\mathbf{A}^T\mathbf{Z}\right)$; and if $\alpha^* > 0$ then $\lambda^* = 0$ which further implies that $\alpha^* = \mathrm{Tr}\left(\mathbf{A}^T\mathbf{Z}\right)$. These conditions imply that optimal α^* is given by

$$\alpha^* = \max\{0, \mathrm{Tr}\left(\mathbf{A}^T\mathbf{Z}\right)\} \tag{2.34}$$

Substituting α^* in the objective function we have

$$\min_{\alpha\geq 0} \alpha^2 - 2\alpha\mathrm{Tr}\left(\mathbf{A}^T\mathbf{Z}\right) = -\left[\max\{0, \mathrm{Tr}\left(\mathbf{A}^T\mathbf{Z}\right)\}\right]^2. \tag{2.35}$$

Using the above result, the error can be stated as

$$\mathcal{E}_{\mathcal{P}C_{m,m',n}}(\mathbf{A}) = \|\mathbf{A}\|_F^2 - \max_{\mathbf{Z}\in\widetilde{\mathcal{P}C}_{m,m',n}} \left[\max\{0, \mathrm{Tr}\left(\mathbf{A}^T\mathbf{Z}\right)\}\right]^2 \tag{2.36}$$

The second term in the above expression can be simplified as

$$\max_{\mathbf{Z}\in\widetilde{\mathcal{P}C}_{m,m',n}} \left[\max\{0, \mathrm{Tr}\left(\mathbf{A}^T\mathbf{Z}\right)\}\right]^2 = \max_{\mathbf{Z}\in\widetilde{\mathcal{P}C}_{m,m',n}} |\mathrm{Tr}\left(\mathbf{A}^T\mathbf{Z}\right)|^2 \tag{2.37}$$

The above equality is true because $\max\{0, \mathrm{Tr}\left(\mathbf{A}^T\mathbf{Z}\right)\} \leq |\mathrm{Tr}\left(\mathbf{A}^T\mathbf{Z}\right)|$ and the set $\widetilde{\mathcal{P}C}_{m,m',n}$ is symmetric with respect to the origin, i.e. if $\mathbf{Z} \in \widetilde{\mathcal{P}C}_{m,m',n}$ then $-\mathbf{Z} \in \widetilde{\mathcal{P}C}_{m,m',n}$, so we always have $\max\{0, \mathrm{Tr}\left(\mathbf{A}^T\mathbf{Z}\right)\} = |\mathrm{Tr}\left(\mathbf{A}^T\mathbf{Z}\right)|$ either at $\mathbf{Z}$ or $-\mathbf{Z}$. Further, we also have that

$$\begin{aligned}\max_{\mathbf{Z}\in\widetilde{\mathcal{P}C}_{m,m',n}} |\mathrm{Tr}\left(\mathbf{A}^T\mathbf{Z}\right)|^2 &= \left[\max_{\mathbf{Z}\in\widetilde{\mathcal{P}C}_{m,m',n}} |\mathrm{Tr}\left(\mathbf{A}^T\mathbf{Z}\right)|\right]^2 \\ &= \left[\max_{\mathbf{Z}\in\widetilde{\mathcal{P}C}_{m,m',n}} \mathrm{Tr}\left(\mathbf{A}^T\mathbf{Z}\right)\right]^2, \end{aligned} \tag{2.38}$$

where the first equality is true by the monotonicity the function $f(t) = t^2$ for $t \geq 0$, and the second equality is again due to the fact that both $\mathbf{Z}$ and $-\mathbf{Z}$ lie in $\widetilde{\mathcal{P}C}_{m,m',n}$. Finally, using (2.37) and (2.38), the error in (2.36) can be written as

$$\mathcal{E}_{\mathcal{P}C_{m,m',n}}(\mathbf{A}) = \|\mathbf{A}\|_F^2 - \left[\max_{\mathbf{Z}\in\widetilde{\mathcal{P}C}_{m,m',n}} \mathrm{Tr}\left(\mathbf{A}^T\mathbf{Z}\right)\right]^2$$

□

2.11.4 Proof of theorem 2.6.1

Proof. The expected value of $\mathcal{E}_{\mathcal{P}C_{m,m',n}}(\mathbf{A})$ is given by

$$\begin{aligned}
&\mathbb{E}_{\mathbf{A}}\left(\mathcal{E}_{\mathcal{P}C_{m,m',n}}(\mathbf{A})\right)\\
&= \mathbb{E}_{\mathbf{A}}\left(\|\mathbf{A}\|_F^2\right) - \mathbb{E}_{\mathbf{A}}\left[\left(\max_{\mathbf{Z}\in\widetilde{\mathcal{P}C}_{m,m',n}} \operatorname{Tr}\left(\mathbf{A}^T\mathbf{Z}\right)\right)^2\right]\\
&= mn - \mathbb{E}_{\mathbf{A}}\left[\left(\max_{\mathbf{Z}\in\widetilde{\mathcal{P}C}_{m,m',n}} \operatorname{Tr}\left(\mathbf{A}^T\mathbf{Z}\right)\right)^2\right],\\
&\leq mn - \left[\mathbb{E}_{\mathbf{A}}\left(\max_{\mathbf{Z}\in\widetilde{\mathcal{P}C}_{m,m',n}} \operatorname{Tr}\left(\mathbf{A}^T\mathbf{Z}\right)\right)\right]^2,
\end{aligned}$$

where the inequality in the last step is due to Jensen's inequality and convexity of the quadratic function. The quantity $\mathbb{E}_{\mathbf{A}}\left(\max_{\mathbf{Z}\in\widetilde{\mathcal{P}C}_{m,m',n}} \operatorname{Tr}\left(\mathbf{A}^T\mathbf{Z}\right)\right)$ is a well known quantity known as Gaussian width we denote it by $\omega(\widetilde{\mathcal{P}C}_{m,m',n})$ which leads us to

$$\mathbb{E}_{\mathbf{A}}\left(\mathcal{E}_{\mathcal{P}C_{m,m',n}}(\mathbf{A})\right) \leq mn - \left[\omega(\widetilde{\mathcal{P}C}_{m,m',n})\right]^2.$$

The Gaussian width of any set $\mathcal{X}$ in $\mathbb{R}^n$ can be related to measure of its size as follows

$$\begin{aligned}
\omega(\mathcal{X}) &= \frac{\gamma_n}{2}\int_{\mathbb{S}^{n-1}}\left(\max_{\mathbf{x}\in\mathcal{X}}\mathbf{x}^T\mathbf{a} - \min_{\mathbf{x}\in\mathcal{X}}\mathbf{x}^T\mathbf{a}\right)d\mathbf{a},\\
&= \frac{\gamma_n}{2} b\left(\mathcal{X}\right),
\end{aligned}$$

where $\mathbb{S}^{n-1}$ denotes the unit sphere in $\mathbb{R}^n$, γ_n is the expected length of a Gaussian random vector in $\mathbb{R}^n$, and $b\left(\mathcal{X}\right)$ is the Gaussian mean width of the set [66]. The Gaussian mean width is the average length of $\mathcal{X}$ along the unit vectors in $\mathbb{R}^n$ and it is one of the fundamental intrinsic volumes of a body arising in the area of combinatorial geometry [67]. □

2.11.5 Proof of lemma 2.6.2

Proof. The Gaussian width $\mathbb{E}\left(f(\mathbf{A}, m')\right)$ of set $\widetilde{\mathcal{P}C}_{m,m',n}$ can be written as follows

$$\begin{aligned}
\omega\left(\widetilde{\mathcal{P}C}_{m,m',n}\right) &= \mathbb{E}\left(\max_{\mathbf{Z}\in\widetilde{\mathcal{P}C}_{m,m',n}} \operatorname{Tr}\left(\mathbf{A}^T\mathbf{Z}\right)\right) \\
&= \mathbb{E}\left(\max_{\mathbf{P}\in\mathbb{R}^{m\times m'},\mathbf{S}\in\mathcal{S}_{m'}} \max_{\mathbf{C}\in\mathcal{C}_n} \operatorname{Tr}\left(\mathbf{A}^T\frac{\mathbf{PSC}}{\|\mathbf{PSC}\|_F}\right)\right) \\
&\geq \mathbb{E}\left(\max_{\mathbf{P}\in\mathbb{R}^{m\times m'},\mathbf{S}\in\mathcal{S}_{m'}} \operatorname{Tr}\left(\mathbf{A}^T\frac{\mathbf{PS}}{\|\mathbf{PS}\|_F}\right)\right).
\end{aligned}$$

where the last step is obtained by fixing $\mathbf{C}$ as $n \times n$ identity matrix. The trace term can be expressed as

$$\begin{aligned}
\operatorname{Tr}\left(\mathbf{A}^T\frac{\mathbf{PS}}{\|\mathbf{PS}\|_F}\right) &= \operatorname{Tr}\left(\left(\mathbf{AS}^T\right)^T\frac{\mathbf{P}}{\|\mathbf{PS}\|_F}\right) \\
&= \operatorname{Tr}\left(\left(\mathbf{AS}^T\right)^T\frac{\mathbf{P}}{\|\mathbf{P}\|_F}\right),
\end{aligned}$$

where the first step utilizes the cyclic property of traces and last step is due to the fact that $\mathbf{S}$ just permutes the columns of $\mathbf{P}$ so that the Frobenius norm of $\mathbf{PS}$ is same as that of $\mathbf{P}$. Next using Cauchy Schwartz's inequality we have

$$\max_{\mathbf{P}\in\mathbb{R}^{m\times m'}} \operatorname{Tr}\left(\left(\mathbf{AS}^T\right)^T\frac{\mathbf{P}}{\|\mathbf{P}\|_F}\right) = \|\mathbf{AS}^T\|_F. \tag{2.39}$$

With this the Gaussian width can be lower bounded as

$$\begin{aligned}
\omega\left(\widetilde{\mathcal{P}C}_{m,m',n}\right) &\geq \mathbb{E}\left(\max_{\mathbf{S}\in\mathcal{S}_{m'}} \|\mathbf{AS}^T\|_F\right) \\
&\geq \mathbb{E}\left(\sqrt{\sum_{i=1}^{m'}\|\mathbf{A}_{:,i}\|_2^2}\right)
\end{aligned}$$

where the last inequality is obtained by fixing $\mathbf{S}$ as first m' rows of $n \times n$ identity matrix. Further, $\mathbb{E}\left(\sqrt{\sum_{i=1}^{m'}\|\mathbf{A}_{:,i}\|_2^2}\right)$ is the expected length of mm'-dimensional standard Gaussian random vector. It is known to be lower bounded as [66]

$$\mathbb{E}\left(\sqrt{\sum_{i=1}^{m'}\|\mathbf{A}_{:,i}\|_2^2}\right) \geq \frac{mm'}{\sqrt{1+mm'}}$$

With this we finally have

$$\omega\left(\widetilde{\mathcal{P}C}_{m,m',n}\right) \geq \frac{mm'}{\sqrt{1+mm'}}.$$

□

2.11.6 Proof of lemma 2.7.1

Proof. Our proof technique involves construction of random row-sampled circulant matrices $\Phi \in \mathbb{C}^{m' \times n}$ as follows

$$\Phi = \mathbf{S}\mathbf{F}\mathbf{D}_{\xi}\mathbf{F}^H, \tag{2.40}$$

where $\mathbf{F}$ is $n \times n$ FFT matrix, $\mathbf{S}$ is the row sub-sampling that chooses m' rows from $\mathbf{F}$ uniformly at random, and $\mathbf{D}_{\xi}$ is random diagonal matrix with diagonal chosen uniformly at random from $\{-1,1\}^n$. The random row-sampled circulant matrix Φ constructed in (2.40) approximately preserves the length of fixed set of points $\mathcal{Y}$ with high probability as captured by the following lemma

Lemma 2.11.2. Let Φ be the random $m \times n$ row sub-sampled circulant matrix constructed as described in (2.40). For the set $\mathcal{Y} \subset \mathbb{C}^n$ let $m \geq C\epsilon^{-2}\log^4(n)\log\left(\frac{4|\mathcal{Y}|}{\eta}\right)\log(\rho^{-1})$ then with probability at least $(1-\rho)(1-\eta)$ (where $\rho, \eta \in (0,1)$) we have

$$(1-\epsilon)\|\mathbf{y}\|_2^2 \leq \|\Phi\mathbf{y}\|_2^2 \leq (1+\epsilon)\|\mathbf{y}\|_2^2,\ \forall \mathbf{y} \in \mathcal{Y}. \tag{2.41}$$

Proof. The proof is provided in section 2.11.7. □

The proof also requires the following lemma on embedding of the inner product.

Lemma 2.11.3. Let $\mathcal{U}$ and $\mathcal{V}$ be set of points in $\mathbb{R}^n$ and suppose that $\Phi \in \mathbb{C}^{m \times n}$ satisfies the following

$$(1-\epsilon)\|\mathbf{y}\|_2^2 \leq \|\Phi\mathbf{y}\|_2^2 \leq (1+\epsilon)\|\mathbf{y}\|_2^2,\ \forall \mathbf{y} \in \mathcal{Y}, \tag{2.42}$$

where $\mathcal{Y} = \left\{\mathbf{u} - \mathbf{v} \middle| \mathbf{u} \in \mathcal{U}, \mathbf{v} \in \{\mathcal{V} \cup -\mathcal{V} \cup i\mathcal{V} \cup -i\mathcal{V}\}\right\}$ then for any $\mathbf{u} \in \mathcal{U}$ and $\mathbf{v} \in \mathcal{V}$ we have

$$\left|\langle \Phi\mathbf{u}, \Phi\mathbf{v}\rangle - \langle \mathbf{u}, \mathbf{v}\rangle\right| \leq \sqrt{2}\epsilon\|\mathbf{u}\|_2\|\mathbf{v}\|_2, \tag{2.43}$$

where the inner product is defined as $\langle \mathbf{u}, \mathbf{v}\rangle = \mathbf{u}^H\mathbf{v}$.

Proof. The proof is provided in section 2.11.8. □

We proceed by fixing $\mathcal{U} = \{\mathbf{A}_{1,:}, \cdots, \mathbf{A}_{m,:}\}$

$$\mathcal{Y} = \{\mathbf{u} - \mathbf{v} | \mathbf{u} \in \mathcal{U}, \mathbf{v} \in \{\mathcal{X} \cup -\mathcal{X} \cup i\mathcal{X} \cup -i\mathcal{X}\}\}$$

and using Lemma 2.11.2 the random row-sampled random circulant matrix $\Phi \in \mathbb{C}^{m' \times n}$ constructed in (2.40) satisfies (2.41) with probability at least $(1-\rho)(1-\eta)$ provided $m' \geq C\epsilon^{-2} \log^4(n) \log\left(\frac{4|\mathcal{Y}|}{\eta}\right) \log(\rho^{-1})$. Further, the error incurred by approximating given $\mathbf{A}$ by matrices of the form $\mathbf{A}\Phi^H\Phi$

$$\|\mathbf{A}\mathbf{x} - \mathbf{A}\Phi^H\Phi\mathbf{x}\|_2^2 = \sum_{i=1}^{m} |\langle \mathbf{A}_{i,:}, \mathbf{x}\rangle - \langle \Phi\mathbf{A}_{i,:}, \Phi\mathbf{x}\rangle|^2,$$

Using Lemma 2.11.3 we have

$$|\langle \mathbf{A}_{i,:}, \mathbf{x}\rangle - \langle \Phi\mathbf{A}_{i,:}, \Phi\mathbf{x}\rangle| \leq \sqrt{2}\epsilon \|\mathbf{A}_{i,:}\|_2$$

which implies that $\|\mathbf{A}\mathbf{x} - \mathbf{A}\Phi^H\Phi\mathbf{x}\|_2$ can be bounded as

$$\|\mathbf{A}\mathbf{x} - \mathbf{A}\Phi^H\Phi\mathbf{x}\|_2 \leq \sqrt{2}\epsilon \|\mathbf{A}\|_F.$$

Further, noticing that $|\mathcal{Y}| = 4|\mathcal{X}|m$ and setting $\rho, \eta = 1/2$ and rescaling $\epsilon := \epsilon/\sqrt{2}$ the number of measurement required are $m' \geq 2C\epsilon^{-2} \log^4(n) \log(32m|\mathcal{X}|) \log(2)$. So finally we can say that there exists a partial circulant matrix Φ from which we can construct $\mathbf{P} = \mathbf{A}\Phi^H$ such we have

$$\|\mathbf{A}\mathbf{x} - \mathbf{P}\Phi\mathbf{x}\|_2 \leq \epsilon \|\mathbf{A}\|_F, \ \forall \mathbf{x} \in \mathcal{X},$$

provided $m' \geq c_1\epsilon^{-2} \log^4(n) \log(c_2 m|\mathcal{X}|)$ where $c_1 = 2C\log(2), c_2 = 32$.

2.11.7 Proof of lemma 2.11.2

The proof uses the following fundamental result on converting a RIP satisfying matrix to Johnson Lindenstrauss (JL) embedding matrix.

Lemma 2.11.4. *[56]* Fix $\eta > 0$ and $\epsilon \in (0,1)$, and consider a finite set $\mathcal{Z} \subset \mathbb{C}^n$ of cardinality $|\mathcal{Z}|$. Suppose the matrix $\Psi \in \mathbb{C}^{m' \times n}$ satisfies the restricted isometry property (RIP) of order (k, δ), i.e., for every $\mathbf{z} \in \mathbb{C}^n$ with at most k non-zeros we have

$(1-\delta)\|\mathbf{z}\|_2^2 \leq \|\Psi\mathbf{z}\|_2^2 \leq (1+\epsilon)\|\mathbf{z}\|_2^2$. Set $k \geq 40\log\left(\frac{4|\mathcal{Z}|}{\eta}\right)$ and $\delta \leq \frac{\epsilon}{4}$. Let $\xi \in \mathbb{R}^n$ be a Rademacher sequence, i.e., uniformly distributed on $\{-1,1\}^n$. Then with probability exceeding $1-\eta$,

$$(1-\epsilon)\|\mathbf{z}\|_2^2 \leq \|\Psi\mathbf{D}_\xi\mathbf{z}\|_2^2 \leq (1+\epsilon)\|\mathbf{z}\|_2^2,\ \forall\mathbf{z} \in \mathcal{Z}. \tag{2.44}$$

We note that this lemma is for complex RIP matrices and vectors. It can be proved by a slight modification of the proof outlined in [56]. Specifically, the main proof is the same, however, one just needs propositions used in the proof to hold for complex numbers as well. It can be easily shown that the propositions are true for complex cases as well. Next from [54] we know that the sub-sampled Fourier matrix **SF**, where **F** is the FFT matrix and **S** is the row sub-sampling matrix that chooses m' rows of **F** uniformly at random, satisfies RIP of level (k,δ) with probability at least $1-\rho$ if $m' \geq C_1 k\delta^{-2}\log^4(n)\log(\rho^{-1})$. Utilizing this fact we invoke Lemma 2.11.4 to convert this matrix satisfying RIP of level (k,δ) to a JL embedding matrix which requires $k \geq 40\log\left(\frac{4|\mathcal{Z}|}{\eta}\right)$ and $\delta \leq \frac{\epsilon}{4}$. This implies that the matrix, $\Psi = \mathbf{SF}$ where **S** is such that the number of rows $m' \geq 640C_1\epsilon^{-2}\log\left(\frac{4|\mathcal{Z}|}{\eta}\right)\log^4(n)\log(\rho^{-1})$ with probability $(1-\rho)(1-\eta)$, we have,

$$(1-\epsilon)\|\mathbf{z}\|_2^2 \leq \|\mathbf{SFD}_\xi\mathbf{z}\|_2^2 \leq (1+\epsilon)\|\mathbf{z}\|_2^2,\ \forall\mathbf{z} \in \mathcal{Z}$$

Further, due to the orthonormality of matrix $\mathbf{F}^H$ in the above inequality we have that,

$$(1-\epsilon)\|\mathbf{Fz}\|_2^2 \leq \|\mathbf{SFD}_\xi\mathbf{F}^H\mathbf{Fz}\|_2^2 \leq (1+\epsilon)\|\mathbf{Fz}\|_2^2,\ \forall\mathbf{z} \in \mathcal{Z}.$$

Choosing $\mathcal{Y} = \{\mathbf{Fz} \mid \mathbf{z} \in \mathcal{Z}\}$ the above inequality implies

$$(1-\epsilon)\|\mathbf{y}\|_2^2 \leq \|\mathbf{SFD}_\xi\mathbf{F}^H\mathbf{y}\|_2^2 \leq (1+\epsilon)\|\mathbf{y}\|_2^2,\ \forall\mathbf{y} \in \mathcal{Y}$$

2.11.8 Proof of lemma 2.11.3

This lemma follows the proof technique of Theorem 4 in [68]. The proof presented in [68] holds for real embedding matrix matrices. Here we extended it to complex embedding matrices.

Consider any $\mathbf{v} \in \mathcal{V}$ and $\mathbf{u} \in \mathcal{U}$. To begin with assume that $\|\mathbf{v}\|_2 = \|\mathbf{u}\|_2 = 1$ and later we will relax this assumption. Using the parallelogram law for $\mathbf{u} \in \mathcal{U}$ and $\mathbf{v} \in \mathcal{V}$

we have

$$\operatorname{Re}\left(\langle \Phi \mathbf{u}, \Phi \mathbf{v} \rangle\right) = \frac{\|\Phi(\mathbf{u}+\mathbf{v})\|_2^2 - \|\Phi(\mathbf{u}-\mathbf{v})\|_2^2}{4}$$

Now using the assumption in (2.42) for $\mathbf{u}+\mathbf{v}$ and $\mathbf{u}-\mathbf{v}$ we have

$$\begin{aligned} &\frac{(1-\epsilon)\|\mathbf{u}+\mathbf{v}\|_2^2 - (1+\epsilon)\|\mathbf{u}-\mathbf{v}\|_2^2}{4} \\ &\qquad \leq \operatorname{Re}\left(\langle \Phi \mathbf{u}, \Phi \mathbf{v} \rangle\right) \leq \frac{(1+\epsilon)\|\mathbf{u}+\mathbf{v}\|_2^2 - (1-\epsilon)\|\mathbf{u}-\mathbf{v}\|_2^2}{4} \end{aligned}$$

Further expanding the terms $\|\mathbf{u}+\mathbf{v}\|_2^2$ and $\|\mathbf{u}-\mathbf{v}\|_2^2$ and re-arranging the inequality we have

$$\left|\operatorname{Re}\left(\langle \Phi \mathbf{u}, \Phi \mathbf{v} \rangle\right) - \langle \mathbf{u}, \mathbf{v} \rangle\right| \leq \epsilon \tag{2.45}$$

Similarly, using the following parallelogram law we have

$$\operatorname{Img}\left(\Phi\langle \mathbf{u}, \Phi \mathbf{v} \rangle\right) = \frac{\|\Phi(\mathbf{u}-i\mathbf{v})\|_2^2 - \|\Phi(\mathbf{u}+i\mathbf{v})\|_2^2}{4}$$

Using the assumption in (2.42) for $\mathbf{u}+i\mathbf{v}$ and $\mathbf{u}-i\mathbf{v}$ we have

$$\begin{aligned} &\frac{(1-\epsilon)\|\mathbf{u}-i\mathbf{v}\|_2^2 - (1+\epsilon)\|\mathbf{u}+i\mathbf{v}\|_2^2}{4} \\ &\qquad \leq \operatorname{Img}\left(\langle \Phi \mathbf{u}, \Phi \mathbf{v} \rangle\right) \leq \frac{(1+\epsilon)\|\mathbf{u}-i\mathbf{v}\|_2^2 - (1-\epsilon)\|\mathbf{u}+i\mathbf{v}\|_2^2}{4} \end{aligned}$$

Now since $\|\mathbf{u}+i\mathbf{v}\|_2^2 = \|\mathbf{u}-i\mathbf{v}\|_2^2 = 2$ we have the following

$$\left|\operatorname{Img}\left(\langle \Phi \mathbf{u}, \Phi \mathbf{v} \rangle\right)\right| \leq \epsilon \tag{2.46}$$

Next we have that

$$\begin{aligned} \left|\langle \Phi \mathbf{u}, \Phi \mathbf{v} \rangle - \langle \mathbf{u}, \mathbf{v} \rangle\right| &= \sqrt{\left|\operatorname{Re}\left(\langle \Phi \mathbf{u}, \Phi \mathbf{v} \rangle\right) - \langle \mathbf{u}, \mathbf{v} \rangle\right|^2 + \left|\operatorname{Img}\left(\langle \Phi \mathbf{u}, \Phi \mathbf{v} \rangle\right)\right|^2} \\ &\leq \sqrt{2}\epsilon, \end{aligned}$$

where the second inequality is by using (2.45) and (2.46). Next using bi-linearity of the inner product the unit norm assumption on $\mathbf{u}$ and $\mathbf{v}$ can be dropped and we have

$$\left|\langle \Phi \mathbf{u}, \Phi \mathbf{v} \rangle - \langle \mathbf{u}, \Phi \mathbf{v} \rangle\right| \leq \sqrt{2}\epsilon \|\mathbf{u}\|_2 \|\mathbf{v}\|_2.$$

□

2.11.9 Proof of theorem 2.8.1

Proof. For $\mathbf{S}$ update step we need to solve

$$\mathbf{S}^{(t)} = \arg\min_{\mathbf{S}\in\mathcal{S}_m} \|\mathbf{A}\mathbf{X} - \mathbf{S}\mathbf{C}^{(t)}\mathbf{X}\|_F^2.$$

The objective function in the above problem can be expanded as

$$\begin{aligned} \|\mathbf{A}\mathbf{X} - \mathbf{S}\mathbf{C}^{(t)}\mathbf{X}\|_F^2 &= \sum_{i=1}^{m} \|\mathbf{A}_{i,:}\mathbf{X} - \mathbf{S}_{i,:}\mathbf{C}^{(t)}\mathbf{X}\|_2^2 \\ &= \sum_{i=1}^{m}\sum_{j=1}^{n} S_{i,j}\|\mathbf{A}_{i,:}\mathbf{X} - \mathbf{C}_{j,:}^{(t)}\mathbf{X}\|_2^2 \end{aligned}$$

where the second equality is due to the fact that $\mathbf{S}_{i,:}$ is some row of identity matrix so it has just 1 non-zero entry. Using this fact the $\mathbf{S}$ update step can be written as

$$\min_{\mathbf{S}\in\mathcal{S}_m} \sum_{i=1}^{m}\sum_{j=1}^{n} S_{i,j}\|\mathbf{A}_{i,:}\mathbf{X} - \mathbf{C}_{j,:}^{(t)}\mathbf{X}\|_2^2 \tag{2.47}$$

Above problem can be transformed to a standard problem known as *linear sum assignment problem* by introducing the constants

$$W_{i,j} = \begin{cases} \|\mathbf{A}_{i,:}\mathbf{X} - \mathbf{C}_{j,:}^{(t)}\mathbf{X}\|_2^2 & 1 \le i \le m, \forall j \\ 0 & i > m, \forall j. \end{cases}$$

as follows

$$\min_{\mathbf{P}\in\mathcal{S}_n} \sum_{i=1}^{n}\sum_{j=1}^{n} P_{i,j}W_{i,j}. \tag{2.48}$$

The above *linear sum assignment* problem can be optimally solved by relaxing it to the following linear program [69]

$$\begin{aligned} \min_{\mathbf{P}\in\mathbb{R}^{n\times n}} \quad & \textstyle\sum_{i=1}^{n}\sum_{i=1}^{n} W_{ij}P_{ij} \\ & \textstyle\sum_{i=1}^{n} P_{ij} = 1, \forall j \\ & \textstyle\sum_{j=1}^{n} P_{ij} = 1, \forall i \\ & P_{ij} \ge 0, \forall i, j. \end{aligned}$$

Notice that in above problem the optimization variable is of size $n \times n$. Also, since the objective cost in 2.48 does not depend on rows other than first m rows of $\mathbf{P}$, the problem 2.47 is optimally solved by choosing first m-rows of the optimal solution 2.48 □

Chapter 3

Noisy matrix and tensor completion under sparse factor models

The task of estimating a signal from its noisy and undersampled observations arises in a variety of applications in signal processing and machine learning. At the extreme end of undersampling lies the case of missing observations. The famous example of inference under missing observations is that of *matrix completion* problem in which signal takes the form of a matrix which is observed at a subset of its entries and the task is to obtain an accurate estimate of the entire matrix. In general, the recovery of missing entries is not possible even if single entry is missing in noiseless setting. However, if the matrix being recovered has a *low-dimensional* structure; exact recovery in noiseless setting and accurate recovery in noisy setting is possible. A well explored structure is that of low rank matrices which has been extensively studied in noise as well as noiseless setting [70–76]. The low rank matrix completion problem arises in many applications, including collaborative filtering [77,78], learning and content analytics [79], sensor network localization [80] etc.

Tensors which may be viewed as generalization of matrices from two-way to multiple-way array naturally arise in many applications in the area of signal processing, computer vision, neuroscience, etc [81, 82]. Often in practice tensor data is collected in a noisy environment and suffers from missing observations which naturally leads to the tensor completion problem. Given the success of matrix completion methods, it is no surprise that recently there has been a lot of interest in extending the successes of matrix completion to tensor completion problem. Similar to matrix completion the structure of low-rank have been exploited in for the tensors as well [83–85].

While the existing literature in matrix/tensor completion overwhelmingly centered around the low rank structure the focus of this chapter is on matrix/tensor completion for matrices following the *sparse factor models*. Specifically, for matrices we consider the completion of matrix which can be written as product of two factors one of which is sparse. Such matrices arise in variety of applications ranging from sparse subspace clustering [86–88] to dictionary learning [89–91] among many others. For tensors, we focus on tensors that admit *sparse CP decomposition* by which we mean that one of the canonical polyadic or CANDECOMP/PARAFAC (CP)-factors is sparse (See (3.1) for definition). Tensors admitting such structure arise in variety of applications involving electroencephalography (EEG) data, neuroimaging using functional magnetic resonance imaging (MRI), and many others [92–96].

In this chapter we discuss the sparsity-regularized maximum likelihood estimation based approach for matrix and tensor completion. We provide general estimation error bounds for noisy matrix and tensor completion via sparsity-regularized maximum likelihood estimation for tensors and matrices following the sparse factor models. These bounds are general enough so that they can be instantiated for variety of noise distributions. After going over a few preliminaries and notations used throughout this chapter we first briefly discuss our contributions on matrix completion for sparse factor models followed by a detailed discussion its extension to the tensor setting.

3.1 Preliminaries and notations

We will denote vectors with lower-case letters, matrices using upper-case letters and tensors as underlined upper-case letters (e.g., $\mathbf{v} \in \mathbb{R}^n, \mathbf{A} \in \mathbb{R}^{m \times n}$, and $\underline{\mathbf{X}} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$, respectively). Furthermore, for any vector (or matrix) $\mathbf{v} \in \mathbb{R}^n$ define $\|\mathbf{v}\|_0 = |\{i : v_i \neq 0\}|$ to be the number of non-zero elements of v and $\|\mathbf{v}\|_\infty := \max_i \{|v_i|\}$ to denote maximum absolute of $\mathbf{v}$. Note that $\|\mathbf{A}\|_\infty := \max_{i,j} \{|A_{i,j}|\}$ is *not* the induced norm of the matrix

$\mathbf{A}$. Entry (i, j, k) of tensor $\underline{\mathbf{X}}$ will be denoted by $X_{i,j,k}$. For a tensor $\underline{\mathbf{X}}$ we define its Frobenius norm in analogy with the matrix case as $\|\underline{\mathbf{X}}\|_F^2 = \sum_{i,j,k} X_{i,j,k}^2$ the squared two norm of its vectorization and its maximum absolute entry as $\|\underline{\mathbf{X}}\|_\infty = \max_{i,j,k} |X_{i,j,k}|$. Finally, we define the canonical polyadic or CANDECOMP/PARAFAC (CP) decomposition of a tensor $\underline{\mathbf{X}} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ to be a representation

$$\underline{\mathbf{X}} = \sum_{f=1}^{F} \mathbf{a}_f \circ \mathbf{b}_f \circ \mathbf{c}_f =: [\mathbf{A}, \mathbf{B}, \mathbf{C}], \tag{3.1}$$

where $\mathbf{a}_f, \mathbf{b}_f$, and $\mathbf{c}_f$ are the f^{th} columns of $\mathbf{A}, \mathbf{B}$, and $\mathbf{C}$, respectively, $\mathbf{a}_f \circ \mathbf{b}_f \circ \mathbf{c}_f$ denotes the tensor outer product such that $(\mathbf{a}_f \circ \mathbf{b}_f \circ \mathbf{c}_f)_{i,j,k} = (i^{th}$ entry of $\mathbf{a}_f) \times (j^{th}$ entry of $\mathbf{b}_f) \times (k^{th}$ entry of $\mathbf{c}_f)$, and $[\mathbf{A}, \mathbf{B}, \mathbf{C}]$ is the shorthand notation of $\underline{\mathbf{X}}$ in terms of its CP factors. The parameter F is an upper bound on the *rank* of $\underline{\mathbf{X}}$ (we refer the reader to [82] for a comprehensive overview of tensor decompositions and their uses). For a given tensor $\underline{\mathbf{X}}$ and CP decomposition $[\mathbf{A}, \mathbf{B}, \mathbf{C}]$ define $n_{\max} = \max\{n_1, n_2, n_3, F\}$ as the maximum dimension of its CP factors and number of latent factors.

3.2 Noisy matrix completion for sparse factor models

Formally, the matrix completion problem in noisy settings can be stated as – Let $\mathbf{X}^* \in \mathbb{R}^{n_1 \times n_2}$ denote a matrix whose elements we wish to estimate, and suppose that we observe $\mathbf{X}^*$ at only a subset $\mathcal{S} \subset [n_1] \times [n_2]$ of its locations, where $[n_1] = \{1, 2, \ldots, n_1\}$, obtaining at each $(i, j) \in \mathcal{S}$ a noisy measurement denoted by $Y_{i,j}$. The overall aim is to estimate $\mathbf{X}^*$ given the observations $\{Y_{i,j}\}_{(i,j) \in \mathcal{S}}$. Without further assumptions the matrix completion problem is ill-posed as the value of $\mathbf{X}^*$ at the unobserved locations could be arbitrary. In context of matrix completion problem there has been extensive focus on low-rank structure. Recent works examining the matrix completion for low-rank structured matrices in noiseless setting include [70–73], in noisy setting include [74–76], and quantized setting include [97–99]. However there are many other low dimensional structures of matrices which usually arise in practice. One specific structure is that of sparse factor model in which the given matrix can be factorized as product of two matrices one of which is sparse. Such matrices arise in many applications including sparse subspace clustering [86–88] and dictionary learning [89–91].

We focus on matrix completion problem for sparse factor models for the matrices with such sparse factorization. Specifically, we consider that the true matrix $\mathbf{X}^* \in \mathbb{R}^{n_1 \times n_2}$ can be factorized as

$$\mathbf{X}^* = \mathbf{D}^* \mathbf{A}^*, \quad \mathbf{D} \in \mathbb{R}^{n_1 \times r}, \mathbf{A} \in \mathbb{R}^{r \times n_2}$$

where $\|\mathbf{D}^*\|_{\max} := \max_{i,j} |D_{i,j}| \leq 1$, $\|\mathbf{A}^*\|_{\max} \leq A_{max}$ for a constant $0 < A_{max} < n_1 \vee n_2$, and $\|\mathbf{X}^*\|_{\max} \leq X_{max}/2$ for a constant $X_{max} \geq 1$. We assume that this true matrix $\mathbf{X}^*$ is observed only at a subset of $\mathcal{S} \subset [n_1] \times [n_2]$ of its entries such that each (i,j) is observed independently and identically with probability $\gamma = m(n_1 n_2)^{-1}$ to obtain the observations Y_{ij}, where m is the nominal number of observations. The observations $\{Y_{i,j\,(i,j)\in\mathcal{S}}\} := \mathbf{Y}_{\mathcal{S}}$ are assumed to be conditionally independent given $\mathcal{S}$ and can be modeled via joint density

$$p_{\mathbf{X}^*_{\mathcal{S}}}(\mathbf{Y}_{\mathcal{S}}) = \prod_{(i,j)\in\mathcal{S}} p_{X^*_{i,j}}(Y_{i,j}) \tag{3.2}$$

Following a sparsity regularized maximum likelihood approach to obtain estimates as follows

$$\hat{\mathbf{X}} = \arg \min_{X = \mathbf{D}\mathbf{A} \in \mathcal{X}} \left\{ -\log p_{\mathbf{X}_{\mathcal{S}}} \mathbf{Y}_{\mathcal{S}} + \lambda \|\mathbf{A}\|_0 \right\}, \tag{3.3}$$

where $\lambda > 0$ is the regularization parameter, $\mathcal{X} \subset \left\{ \mathbf{D}\mathbf{A} \middle| \mathbf{D} \in \mathcal{D}, \mathbf{A} \in \mathcal{A}, \|\mathbf{X}\|_{\max} \leq X_{\max} \right\}$, $\mathcal{D}$ is the set of matrices $\mathbf{D} \in \mathbb{R}^{n_1 \times r}$ whose elements are discretized to one of $L = 2^{\lceil \log(\max n_1, n_2) \rceil^{\beta}}$ (for some fixed $\beta \geq 1$) uniformly spaced values in the range $[-1, 1]$, and $\mathcal{A}$ is the set of matrices $\mathbf{D} \in \mathbb{R}^{r \times n_2}$ whose elements either take the value zero, or are discretized to one of L uniformly spaced values in the range $[-A_{\max}, A_{\max}]$. The main advantage of sparsity regularized maximum likelihood approach in (3.3) is its generality as it can handle various noise probability densities and even non-linear observations setups such as 1-bit quantized observations. The estimate $\hat{X}$ obtained from (3.3) can be shown to satisfy the following general error bound [2]

Theorem 3.2.1. *Let the sample set $\mathcal{S}$ be drawn from the independent Bernoulli model with $\gamma = m(n_1 n_2)^{-1}$ as described above, and let $\mathbf{Y}_{\mathcal{S}}$ be described by* (3.2)*. If C_D is any constant satisfying*

$$C_D \geq \max_{\mathbf{X} \in \mathcal{X}} \max_{i,j} \; D(p_{X^*_{i,j}} \| p_{X_{i,j}}), \tag{3.4}$$

where $\mathcal{X}$ is as above for some $\beta \geq 1$, then for any

$$\lambda \geq 2 \cdot (\beta + 2) \cdot \left(1 + \frac{2C_D}{3}\right) \cdot \log(n_1 \vee n_2), \tag{3.5}$$

the complexity penalized maximum likelihood estimator (3.3) *satisfies the (normalized, per-element) error bound*

$$\begin{aligned} \frac{\mathbb{E}_{\mathcal{S},\mathbf{Y}_\mathcal{S}}\left[-2\log H(p_{\widehat{\mathbf{X}}}, p_{\mathbf{X}^*})\right]}{n_1 n_2} &\leq \frac{8C_D \log m}{m} \\ + \quad 3 \cdot \min_{\mathbf{X}\in\mathcal{X}} &\left\{ \frac{D(p_{\mathbf{X}^*} \| p_{\mathbf{X}})}{n_1 n_2} + \left(\lambda + \frac{4C_D(\beta+2)\log(n_1 \vee n_2)}{3}\right)\left(\frac{n_1 p + \|\mathbf{A}\|_0}{m}\right)\right\} \end{aligned} \tag{3.6}$$

where $n_1 \vee n_2 = \max_{n_1,n_2}$, $D(p\|q) = \mathbb{E}_p\left[\log \frac{p(Y)}{q(Y)}\right]$ *is the KL divergence and* $H(p,q) = \mathbb{E}_p\left[\sqrt{\frac{q(Y)}{p(Y)}}\right]$ *is the Hellinger affinity.*

Proof. The proof details can be found in [2]. □

The above oracle bound was used to obtain error bounds for Gaussian, Laplace, Poisson noise distributions and even for 1-bit quantized observation. A scalable alternating directions method of multiplies (ADMM) based algorithm for solving sparsity regularized maximum likelihood problem was also proposed and empirical justification for the error bounds was also provided under various noise distributions [2]. Recently in [100] the bounds we obtained in [2] were shown to be minimax optimal under linear sparsity regime.

3.3 Noisy tensor completion for tensors with a sparse canonical polyadic factor

We consider the general problem of tensor completion. Let $\underline{\mathbf{X}}^* \in \mathbb{R}^{n_1\times n_2\times n_3}$ be the tensor we wish to estimate and suppose we collect the noisy measurements $Y_{i,j,k}$ at subset of its location $(i,j,k) \in \mathcal{S} \subset [n_1]\times[n_2]\times[n_3]$.[1] The goal of tensor completion problem is to estimate the tensor $\underline{\mathbf{X}}^*$ from noisy observations $\{Y_{i,j,k}\}_{(i,j,k)\in\mathcal{S}}$. This problem is naturally ill-posed without any further assumption on the tensor we wish to

[1] The material in this section is ©2017 IEEE. Reprinted, with permission, from *IEEE International Symposium on Information Theory*, "Noisy tensor completion for tensors with a sparse canonical polyadic factor", S. Jain, A. Gutierrez, and J. Haupt.

estimate. A common theme in recent tensor completion works is to use the tools that have been effective in tackling the matrix completion problem and apply them to tensor completion problem. For example, one could apply matrix completion results to tensors directly by matricizing the tensors along various modes and minimizing the sum or weighted sum of their nuclear norms as a convex proxy for the tensor rank [83–85]. Since the nuclear norm is computationally intractable for large scale data, matrix completion via alternating minimization was extended to tensors in [101, 102]. In contrast to these works, here we focus on structured tensors that admit *"sparse CP decomposition"* by which we mean that one of the canonical polyadic or CANDECOMP/PARAFAC (CP)-factors is sparse.

Recently, the completion of tensors with this model was exploited in the context of time series prediction of incomplete EEG data [94]. Here we focus on providing recovery guarantees and a general algorithmic framework and extend the results of noisy matrix completion under sparse factor model in [2] to tensors with a sparse CP factor.

3.3.1 Data model

Let $\underline{\mathbf{X}}^* \in \mathcal{X} \subset \mathbb{R}^{n_1 \times n_2 \times n_3}$ be the unknown tensor whose entries we wish to estimate. We assume that X^* admits a CP decomposition such that the CP factors $\mathbf{A}^* \in \mathbb{R}^{n_1 \times F}$, $\mathbf{B}^* \in \mathbb{R}^{n_2 \times F}$, $\mathbf{C}^* \in \mathbb{R}^{n_3 \times F}$ are entry-wise bounded: $\|\mathbf{A}^*\|_\infty \leq A_{\max}$, $\|\mathbf{B}^*\|_\infty \leq B_{\max}$, $\|\mathbf{C}^*\|_\infty \leq C_{\max}$. Furthermore, we will assume that $\mathbf{C}^*$ is sparse $\|\mathbf{C}^*\|_0 \leq k$. Then $\underline{\mathbf{X}}^*$ can be decomposed as follows

$$\underline{\mathbf{X}}^* = [\mathbf{A}^*, \mathbf{B}^*, \mathbf{C}^*] = \sum_{f=1}^{F} \mathbf{a}_f^* \circ \mathbf{b}_f^* \circ \mathbf{c}_f^*.$$

$\underline{\mathbf{X}}$ is also entry-wise bounded, say by $\|\underline{\mathbf{X}}^*\|_\infty \leq \frac{X_{\max}}{2}$[2] .Such tensors have a rank upper bounded by F.

3.3.2 Observation setup

We assume that we measure a noisy version of $\underline{\mathbf{X}}^*$ at some random subset of the entries $\mathcal{S} \subset [n_1] \times [n_2] \times [n_3]$. We generate S via an independent Bernoulli model with parameter

[2] Here, the factor of 1/2 is chosen to facilitate the exposition of proof. Any factor in (0, 1) would suffice.

$\gamma \in (0,1]$ as follows: first generate $n_1 n_2 n_3$ i.i.d. Bernoulli random variables $b_{i,j,k}$ with $\text{Prob}(b_{i,j,k} = 1) = \gamma, \forall i,j,k$ and then the set $\mathcal{S}$ is obtained as $\mathcal{S} = \{(i,j,k) : b_{i,j,k} = 1\}$. Conditioned on $\mathcal{S}$, in the case of an additive noise model we obtain noisy observations at the locations of $\mathcal{S}$ as follows

$$Y_{i,j,k} = X^*_{i,j,k} + n_{i,j,k}, \quad \forall (i,j,k) \in \mathcal{S}, \tag{3.7}$$

where $n_{i,j,k}$'s are the i.i.d noise entries.

3.3.3 Estimation procedure

Our goal here is to obtain an estimate for full true tensor $\underline{\mathbf{X}}^*$ using the noisy sub-sampled measurement $\{Y_{i,j,k}\}_{i,j,k \in \mathcal{S}}$. We pursue the sparsity-regularized maximum likelihood to achieve this goal. For this we first note that the observations $Y_{i,j,k}$ have distribution parameterized by the entries of the true tensor $\underline{\mathbf{X}}^*$ and the overall likelihood is given by

$$p_{X^*_{\mathcal{S}}}(\underline{\mathbf{Y}}_{\mathcal{S}}) := \prod_{(i,j,k) \in \mathcal{S}} p_{X^*_{i,j,k}}(Y_{i,j,k}). \tag{3.8}$$

where $p_{X^*_{i,j,k}}(Y_{i,j,k})$ is the pdf of observation $Y_{i,j,k}$ which depends on the pdf of the noise and is parametrized by $X^*_{i,j,k}$, and we have used the shorthand notation $\underline{\mathbf{X}}_{\mathcal{S}}$ to denote the entries of the tensor $\underline{\mathbf{X}}$ sampled at the indices in $\mathcal{S}$.

Using prior information that $\mathbf{C}$ is sparse, we regularize with respect to the sparsity of $\mathbf{C}$ and obtain the sparsity-regularized maximum likelihood estimate $\hat{\underline{\mathbf{X}}}$ of $\underline{\mathbf{X}}^*$ as given below

$$\hat{\underline{\mathbf{X}}} = \underset{\underline{\mathbf{X}} = [\mathbf{A},\mathbf{B},\mathbf{C}] \in \mathcal{X}}{\arg\min} \left(-\log p_{\underline{\mathbf{X}}_{\mathcal{S}}}(\underline{\mathbf{Y}}_{\mathcal{S}}) + \lambda \|\mathbf{C}\|_0 \right), \tag{3.9}$$

where $\lambda > 0$ is the regularization parameter and $\mathcal{X}$ is a class of candidate estimates. Specifically, we take $\mathcal{X}$ to be a finite class of estimates constructed as follows: first choose some $\beta \geq 1$, and set $L_{\text{lev}} = 2^{\lceil \log_2 (n_{max})^\beta \rceil}$ and construct $\mathcal{A}$ to be the set of all matrices $\mathbf{A} \in \mathbb{R}^{n_1 \times F}$ whose elements are discretized to one of L_{lev} uniformly spaced between $[-A_{\max}, A_{\max}]$, similarly construct $\mathcal{B}$ to be the set of all matrices $\mathbf{B} \in \mathbb{R}^{n_2 \times F}$ whose elements are discretized to one of L_{lev} uniformly spaced between $[-B_{\max}, B_{\max}]$, finally $\mathcal{C}$ be the set of matrices $\mathbf{C} \in \mathbb{R}^{n_3 \times F}$ whose elements are either zero or are discretized to one of L_{lev} uniformly spaced between $[-C_{\max}, C_{\max}]$. Then, we let

$$\mathcal{X}' = \left\{ [\mathbf{A}, \mathbf{B}, \mathbf{C}] \middle| \mathbf{A} \in \mathcal{A}, \mathbf{B} \in \mathcal{B}, \mathbf{C} \in \mathcal{C}, \|\underline{\mathbf{X}}\|_\infty \leq X_{\max} \right\} \tag{3.10}$$

and we let $\mathcal{X}$ be any subset of $\mathcal{X}'$.

3.3.4 General error bound

In this section we present the main result in which we provide an upper bound on the quality of the estimate obtained by solving (3.9).

Theorem 3.3.1. *Let $\mathcal{S}$ be sampled according to the independent Bernoulli model with parameter $\gamma = \frac{m}{n_1 n_2 n_3}$ and let $Y_{\mathcal{S}}$ be given by* (3.8)*. Let Q_D be any upper bound on the maximum KL divergence between $p_{X^*_{i,j,k}}$ and $p_{X_{i,j,k}}$ for $\underline{\mathbf{X}} \in \mathcal{X}$*

$$Q_D \geq \max_{\underline{\mathbf{X}} \in \mathcal{X}} \max_{i,j,k} D\left(p_{X^*_{i,j,k}} \| p_{X_{i,j,k}}\right)$$

where $\mathcal{X} \subseteq \mathcal{X}'$ with $\mathcal{X}'$ as defined in (3.10)*. Then for any λ satisfying*

$$\lambda \geq 4\,(\beta + 2)\left(1 + \frac{2Q}{3}\right) \log n_{max} \tag{3.11}$$

the regularized constrained maximum likelihood estimate $\hat{\underline{\mathbf{X}}}$ obtained from (3.9) *satisfies*

$$\begin{aligned} &\frac{\mathbb{E}_{\mathcal{S}, Y_{\mathcal{S}}}\left[-2\log(H(p_{\hat{\underline{\mathbf{X}}}}, p_{\underline{\mathbf{X}}^*}))\right]}{n_1 n_2 n_3} \\ &\quad \leq 3 \min_{\underline{\mathbf{X}} \in \mathcal{X}} \left\{ \frac{D(p_{\underline{\mathbf{X}}^*} \| p_{\underline{\mathbf{X}}})}{n_1 n_2 n_3} + \left(\lambda + \frac{8 Q_D (\beta + 2) \log n_{max}}{3}\right) \right. \\ &\qquad \left. \frac{(n_1 + n_2)F + \|\mathbf{C}\|_0}{m} \right\} + \frac{8 Q_D \log m}{m}. \end{aligned} \tag{3.12}$$

Proof. The proof appears in the section 3.5.1. □

The above theorem extends the main result of [2] to the tensor case. It states a general result relating the log affinity between the distributions parameterized by the estimated tensor and the ground truth tensor. Hellinger affinity is a measure of distance between two probability distributions which can be used to get bounds on the quality of the estimate. As in [2], the main utility of this theorem is that it can be instantiated for noise distributions of interest such as Gaussian, Laplace and Poisson. Note that since the estimation procedure depends only on the likelihood term, the above theorem can also be extended to non-linear observation models such as 1-bit quantized measurements. We next demonstrate the utility of the above theorem to present error guarantees when the additive noise follows a Gaussian distribution.

3.3.5 Implication for Gaussian noise

We examine the implications of Theorem 3.3.1 in a setting where observations are corrupted by independent additive zero-mean Gaussian noise with known variance. In this case, the observations $\mathbf{Y}_{\mathcal{S}}$ are distributed according to a multivariate Gaussian density of dimension $|\mathcal{S}|$ whose mean corresponds to the tensor entries at the sample locations and with covariance matrix $\sigma^2\mathbf{I}_{|\mathcal{S}|}$, where $\mathbf{I}_{|\mathcal{S}|}$ is the identity matrix of dimension $|\mathcal{S}|$. That is,

$$p_{\underline{\mathbf{X}}^*_{\mathcal{S}}}(\underline{\mathbf{Y}}_{\mathcal{S}}) = \frac{1}{(2\pi\sigma^2)^{|\mathcal{S}|/2}} \exp\left(-\frac{1}{2\sigma^2}\ \|\underline{\mathbf{Y}}_{\mathcal{S}} - \underline{\mathbf{X}}^*_{\mathcal{S}}\|_F^2\right), \tag{3.13}$$

In order to apply Theorem 3.3.1 we choose β as:

$$\beta = \max\left\{1, 1 + \frac{\log\left(\frac{14FA_{\max}B_{\max}C_{\max}}{X_{\max}} + 1\right)}{\log(n_{\max})}\right\} \tag{3.14}$$

Then, we fix $\mathcal{X} = \mathcal{X}'$, and obtain an estimate according to (3.9) with the λ value chosen as

$$\lambda = 4\left(1 + \frac{2Q_D}{3}\right)(\beta + 2)\cdot\log(n_{\max}) \tag{3.15}$$

In this setting we have the following result.

Corollary 3.3.1. *Let β be as in* (3.14)*, let λ be as in* (3.15) *with $Q_D = 2X_{\max}^2/\sigma^2$, and let $\mathcal{X} = \mathcal{X}'$. The estimate $\widehat{\underline{\mathbf{X}}}$ obtained via* (3.9) *satisfies*

$$\frac{\mathbb{E}_{\mathcal{S},Y_{\mathcal{S}}}\left[\|\underline{\mathbf{X}}^* - \widehat{\underline{\mathbf{X}}}\|_F^2\right]}{n_1n_2n_3} = \mathcal{O}\left(\log(n_{\max})(\sigma^2 + X_{\max}^2)\left(\frac{(n_1+n_2)F + \|\mathbf{C}^*\|_0}{m}\right)\right). \tag{3.16}$$

Proof. The proof appears in section 3.5.3. □

Remark 1. *The quantity* $(n_1 + n_2)F + \|\mathbf{C}^*\|_0$ *can be viewed as the number of degrees of freedom of the model. In this context, we note that our estimation error is proportional to the number of degrees of freedom of the model divided by* m *multiplied by the logarithmic factor* $\log(n_{\max})$*.*

Remark 2. *If we were to ignore the multilinear structure and matricize the tensor as*

$$\mathbf{X}^*_{(3)} = (\mathbf{B}^* \odot \mathbf{A}^*)(\mathbf{C}^*)^T,$$

where $\odot$ *is the Khatri-Rao product (for details of matricization refer [81]). The matrix* $\mathbf{X}^*_{(3)}$ *follows the sparse factor with* $(\mathbf{C}^*)^T$ *as the sparse factor and* $\mathbf{B}^* \odot \mathbf{A}^*$ *as the dense factor of size* $(n_1 \cdot n_2) \times F$*. Applying corollary III.1 from [2] we would obtain the bound*

$$\frac{\mathbb{E}_{\mathcal{S},Y_{\mathcal{S}}}\left[\|\underline{\mathbf{X}}^* - \widehat{\underline{\mathbf{X}}}\|_F^2\right]}{n_1 n_2 n_3} = \mathcal{O}\left(\log(n_{\max})(\sigma^2 + X_{\max}^2)\left(\frac{(n_1 \cdot n_2)F + \|\mathbf{C}^*\|_0}{m}\right)\right),$$

That is, the factor of $(n_1 + n_2)F$ *in Theorem 3.3.1 has become a factor of* $(n_1 \cdot n_2)F$ *when matricizing, a potentially massive improvement.*

3.3.6 The algorithmic framework

In this section we propose an ADMM-type algorithm to solve the complexity regularized maximum likelihood estimate problem in (3.9). We note that the feasible set $\mathcal{X}$ problem in (3.9) is discrete which makes the algorithm design difficult. Similar to [2] we drop the discrete assumption in order to use continuous optimization techniques. This may be justified by choosing a very large value of L_{lev} and by noting that continuous optimization algorithms, when executed on a computer, use finite precision arithmetic, and thus a discrete set of points. Hence, we consider the design of an optimization algorithm for the following problem:

$$\begin{aligned}
&\min_{\underline{\mathbf{X}},\mathbf{A},\mathbf{B},\mathbf{C}} -\log p_{\underline{\mathbf{X}}_{\mathcal{S}}}(\underline{\mathbf{Y}}_{\mathcal{S}}) + \lambda \|\mathbf{C}\|_0 \\
&\text{subject to} \quad \mathbf{A} \in \mathcal{A}, \mathbf{B} \in \mathcal{B}, \mathbf{C} \in \mathcal{C}, \\
&\qquad \|\underline{\mathbf{X}}\|_\infty \leq X_{\max}, \underline{\mathbf{X}} = \sum_{f=1}^{F} \mathbf{a}_f \circ \mathbf{b}_f \circ \mathbf{c}_f, \\
&\qquad \mathcal{A} = \left\{\mathbf{A} \in \mathbb{R}^{n_1 \times F} : \|\mathbf{A}\|_\infty \leq A_{\max}\right\}, \\
&\qquad \mathcal{B} = \left\{\mathbf{B} \in \mathbb{R}^{n_2 \times F} : \|\mathbf{B}\|_\infty \leq B_{\max}\right\}, \\
&\qquad \mathcal{C} = \left\{\mathbf{C} \in \mathbb{R}^{n_3 \times F} : \|\mathbf{C}\|_\infty \leq C_{\max}\right\}.
\end{aligned} \tag{3.17}$$

We form the augmented Lagrangian for the above problem

$$\begin{aligned}
\mathcal{L}(\underline{\mathbf{X}}, \mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{L}) \quad = \quad & -\log p_{\underline{\mathbf{X}}_{\mathcal{S}}}(\underline{\mathbf{Y}}_{\mathcal{S}}) + \lambda \|\mathbf{C}\|_0 + \frac{\rho}{2} \left\| \underline{\mathbf{X}} - \sum_{f=1}^{F} \mathbf{a}_f \circ \mathbf{b}_f \circ \mathbf{c}_f \right\|_F^2 + \\
& \mathbf{L}^T \cdot \mathbf{vec}\left(\underline{\mathbf{X}} - [\mathbf{A}, \mathbf{B}, \mathbf{C}]\right) + I_{\mathcal{X}}(\underline{\mathbf{X}}) + I_{\mathcal{A}}(\mathbf{A}) + I_{\mathcal{B}}(\mathbf{B}) + I_{\mathcal{C}}(\mathbf{C}),
\end{aligned}$$

where $\mathbf{L}$ is Lagrangian vector of size $n_1 n_2 n_3$ for the tensor equality constraint and $I_{\mathcal{X}}(\underline{\mathbf{X}}), I_{\mathcal{A}}(\mathbf{A}), I_{\mathcal{B}}(\mathbf{B}), I_{\mathcal{C}}(\mathbf{C})$ are indicator functions of the sets $\|\underline{\mathbf{X}}\|_\infty \leq X_{\max}$, $\mathcal{A}$, $\mathcal{B}$, $\mathcal{C}$ respectively[3] . Starting from the augmented Lagrangian we propose the ADMM-type algorithm for the tensor case as shown in Algorithm 4.

[3] The convex indicator of set U is defined as $I_U(x) = $ if $x \in U$ and $I_U(x) = \infty$if $x \notin U$. Note that function $I_U(x)$ is convex function if U is convex set.

Algorithm 4 ADMM-type algorithm for noisy tensor completion

```
**Inputs:** $\Delta_1^{\text{stop}}, \Delta_2^{\text{stop}}, \eta, \rho^{(0)}$
**Initialize:** $\underline{\mathbf{X}}^{(0)}$, $\mathbf{A}^{(0)}$,$\mathbf{B}^{(0)}$,$\mathbf{C}^{(0)}$, $\mathbf{L}^{(0)}$
  **while** $\Delta_1 > \Delta_1^{\text{stop}}, \Delta_2 > \Delta_2^{\text{stop}}, t \leq t_{max}$ **do**
    **S1:** $\underline{\mathbf{X}}^{(t+1)} = \arg\min_{\underline{\mathbf{X}}} \mathcal{L}(\underline{\mathbf{X}}, \mathbf{A}^{(t)}, \mathbf{B}^{(t)}, \mathbf{C}^{(t)}, \mathbf{L}^{(t)})$
    **S2:** $\mathbf{A}^{(t+1)} = \arg\min_{\mathbf{A}} \mathcal{L}(\underline{\mathbf{X}}^{(t+1)}, \mathbf{A}, \mathbf{B}^{(t)}, \mathbf{C}^{(t)}, \mathbf{L}^{(t)})$
    **S3:** $\mathbf{B}^{(t+1)} = \arg\min_{\mathbf{B}} \mathcal{L}(\underline{\mathbf{X}}^{(t+1)}, \mathbf{A}^{(t+1)}, \mathbf{B}, \mathbf{C}^{(t)}, \mathbf{L}^{(t)})$
    **S4:** $\mathbf{C}^{(t+1)} = \arg\min_{\mathbf{C}} \mathcal{L}(\underline{\mathbf{X}}^{(t+1)}, \mathbf{A}^{(t+1)}, \mathbf{B}^{(t+1)}, \mathbf{C}, \mathbf{L}^{(t)})$
    **S5:** $\mathbf{L}^{(t+1)} = \mathbf{L}^{(t)} + \rho^{(0)}\mathbf{vec}\left(\underline{\mathbf{X}}^{(t+1)} - [\mathbf{A}^{(t+1)}, \mathbf{B}^{(t+1)}, \mathbf{C}^{(t+1)}]\right)$
    Set $\Delta_1 = \left\|\underline{\mathbf{X}}^{(t+1)} - [\mathbf{A}^{(t+1)}, \mathbf{B}^{(t+1)}, \mathbf{C}^{(t+1)}]\right\|_F$
    Set $\Delta_2 = \rho^{(k)} \left\|[\mathbf{A}^{(t)}, \mathbf{B}^{(t)}, \mathbf{C}^{(t)}] - [\mathbf{A}^{(t+1)}, \mathbf{B}^{(t+1)}, \mathbf{C}^{(t+1)}]\right\|_F$
    $\rho^{(k+1)} = \begin{cases} \eta\rho^{(k)}, \text{ if} \Delta_1 \geq 10\Delta_2 \\ \rho^{(k)}/\eta, \text{ if} \Delta_2 \geq 10\Delta_1 \\ \rho^{(k)}, \text{ otherwise} \end{cases}$
  **end while**
**Output:** $\mathbf{A} = \mathbf{A}^{(t)}, \mathbf{B} = \mathbf{B}^{(t)}, \mathbf{C} = \mathbf{C}^{(t)}$
```

The $\underline{\mathbf{X}}$ update in Algorithm 4 is separable across components and so it reduces to $n_1 n_2 n_3$ scalar problems. Furthermore, the scalar problem is closed-form for $(i, j, k) \notin \mathcal{S}$ and is a proximal-type step for $(i, j, k) \in \mathcal{S}$. This is a particularly attractive feature because many common noise densities (e.g., Gaussian, Laplace) have closed-form proximal updates. The $\mathbf{A}$ and $\mathbf{B}$ updates can be converted to a constrained least squares problem and can be solved via projected gradient descent. We solve the $\mathbf{C}$ update via iterative hard thresholding. Although the convergence of this algorithm to a stationary point remains an open question and a subject of future work, we have not encountered problems with this in our simulations.

3.3.7 Experimental evaluation

In this section we include simulations which corroborate our theorem. For each experiment we construct the true data tensor $\underline{\mathbf{X}} = [\mathbf{A}^*, \mathbf{B}^*, \mathbf{C}^*]$ by individually constructing the CP factors $\mathbf{A}^*, \mathbf{B}^*, \mathbf{C}^*$ (as described below), where the magnitudes of entries of the true factors $\mathbf{A}^*$, $\mathbf{B}^*$, and $\mathbf{C}^*$ are bounded in magnitude by $A^*_{\max}, B^*_{\max}$, and $C^*_{\max}$ respectively. For the purposes of these experiments we fix $n_1 = 30, n_2 = 30, n_3 = 50$ and $A^*_{\max} = 1, B^*_{\max} = 1, C^*_{\max} = 10$.

For a given F the true CP factors were generated as random matrices of dimensions $n_1 \times F$, $n_2 \times F$, $n_3 \times F$ with standard Gaussian $\mathcal{N}(0,1)$ entries. We then projected the entries of the $\mathbf{A}$ and $\mathbf{B}$ matrices so that $\|\mathbf{A}^*\|_\infty \leq A^*_{\max}$ and $\|\mathbf{B}^*\|_\infty \leq B^*_{\max}$. For the $\mathbf{C}^*$ matrix we first project $\mathbf{C}^*$ entry-wise to the interval $[-C_{\max}, C_{\max}]$ and then pick k entries uniformly at random and zero out all other entries so that we get the desired sparsity $\|\mathbf{C}^*\|_0 = k$. From these tensors the tensor $\underline{\mathbf{X}}^*$ was calculated as $\underline{\mathbf{X}}^* = [\mathbf{A}^*, \mathbf{B}^*, \mathbf{C}^*]$ as in (3.1).

We then take measurements at a subset of entries following a Bernoulli sampling model with sampling rate $\gamma \in (0,1]$ and corrupt our measurements with additive white Gaussian noise of variance $\sigma = 0.25$ to obtain the final noisy measurements. The noisy measurements were then used to calculate the estimate by solving (an approximation to) the complexity regularized problem in (3.17) using algorithm 4. Note that for Gaussian noise the negative log-likelihood in problem (3.17) reduces to a squared error loss over the sampled entries. Since in practice the parameters $A_{\max}$, $B_{\max}$, $C_{\max}, X_{\max}$ are not known a priori we will assume we have an upper bound for them and in our experiments set them as $A_{\max} = 2A^*_{\max}, B_{\max} = 2B^*_{\max}, C_{\max} = 2C^*_{\max}, X_{\max} = 2\|\underline{\mathbf{X}}^*\|_\infty$. Further, we also assume that F is known a priori.

In figure 3.1 we show how the log per entry squared error $\log\left(\frac{\|\hat{\underline{\mathbf{X}}} - \underline{\mathbf{X}}^*\|_F^2}{n_1 n_2 n_3}\right)$ decays as a function of log sampling rate $\log(\gamma)$ for $F = 5, 15$ in the paper and a fixed sparsity level $\|\mathbf{C}\|_0 = 0.2 n_3 F$. The plot is obtained after averaging over 10 trials to average out random Bernoulli sampling at given sampling rate γ and noise. Each plot corresponds to a single chosen value of λ, selected as the value that gives a representative error curve (e.g., one giving lowest overall curve, over the range of parameters we considered). Our theoretical results predict that the error decay should be inversely proportional to the sampling rate $\gamma = \frac{m}{n_1 n_2 n_3}$ when viewed on a log-log scale, this corresponds to the slope of -1. The curve of $F = 5$ and $F = 15$ are shown in blue solid line and red dotted line. For both the cases the slope of curves is similar and it is approximately -1. Therefore these experimental results validate both the theoretical error bound in corollary 3.3.1 and the performance of our proposed algorithm.

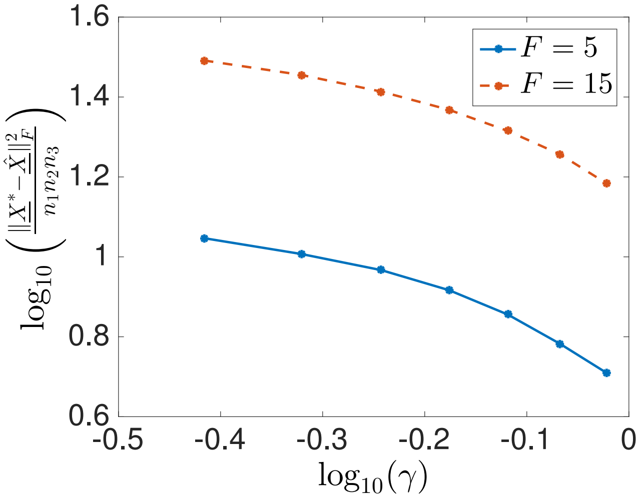


Figure 3.1: Plot for log per-entry approximation error vs the log sampling rate for the two ranks $F = 5, 15$. The slope at the higher sampling rates is approximately -1 (the rate predicted by our theory) in both cases.

3.4 Summary

We consider problem matrix and tensor completion from noisy missing observations for sparse factor models. We proposed a sparsity regularized maximum likelihood approach and provided general performance error guarantees for matrix as well as tensor completion. The utility of the performance bound for tensor completion was demonstrated by instantiating it for the Gaussian noise case. We also provided a ADMM-style algorithm for tensor completion which was used to provide numerical evidence to the theoretical error bounds. See section 7.2 for more discussion on possible future directions of research for noisy tensor completion.

3.5 Appendix

3.5.1 Proof of theorem 3.3.1

The proof of our main result is an application of the following general lemma.

Lemma 3.5.1. *Let $\underline{\mathbf{X}}^* \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ and let $\mathcal{X}$ be a finite collection of candidate reconstructions with assigned weights* $\text{pen}(\underline{\mathbf{X}}) \geq 1$ *satisfying the Kraft-McMillan inequality over $\mathcal{X}$.*

$$\sum_{\underline{\mathbf{X}} \in \mathcal{X}} 2^{-\text{pen}(\underline{\mathbf{X}})} \leq 1. \tag{3.18}$$

Fix an integer $4 \leq m \leq n_1 n_2 n_3$ and let $\gamma = \frac{m}{n_1 n_2 n_3}$ and generate $n_1 n_2 n_3$ i.i.d. Bernoulli(γ) random variables $b_{i,j,k}$ so that entry $(i,j,k) \in \mathcal{S}$ if $b_{i,j,k} = 1$ and $(i,j,k) \notin \mathcal{S}$ otherwise. Conditioned on $\mathcal{S}$ we obtain independent measurements $\underline{\mathbf{Y}}_{\mathcal{S}} \sim p_{\underline{\mathbf{X}}^*_{\mathcal{S}}} = \prod_{(i,j,k) \in \mathcal{S}} p_{X^*_{i,j,k}}$. Then if Q_D is an upper bound for the maximum KL-divergence

$$Q_D \geq \max_{\underline{\mathbf{X}} \in \mathcal{X}} \max_{(i,j,k)} D(p_{X^*_{i,j,k}} \| p_{X_{i,j,k}}),$$

it follows that for any

$$\xi \geq (1 + \frac{2Q_D}{3}) \cdot 2 \log 2 \tag{3.19}$$

the complexity-penalized maximum likelihood estimator

$$\hat{\underline{\mathbf{X}}}^{\xi}(\mathcal{S}, \underline{\mathbf{Y}}_{\mathcal{S}}) = \arg\min_{\underline{\mathbf{X}} \in \mathcal{X}} \left\{ -\log p_{\underline{\mathbf{X}}_{\mathcal{S}}}(\underline{\mathbf{Y}}_{\mathcal{S}}) + \xi \text{pen}(\underline{\mathbf{X}}) \right\}$$

satisfies the error bound

$$\frac{\mathbb{E}_{\mathcal{S}, \underline{\mathbf{Y}}_{\mathcal{S}}} \left[-2 \log(H(p_{\hat{\underline{\mathbf{X}}}^*}, p_{\underline{\mathbf{X}}^*})) \right]}{n_1 n_2 n_3} \leq \frac{8 Q_D \log m}{m} + \\ 3 \min_{\underline{\mathbf{X}} \in \mathcal{X}} \left\{ \frac{D(p_{\underline{\mathbf{X}}^*} \| p_{\underline{\mathbf{X}}})}{n_1 n_2 n_3} + \left(\xi + \frac{4 Q_D \log 2}{3} \right) \frac{\text{pen}(\underline{\mathbf{X}})}{m} \right\}.$$

Proof. The proof appears in Appendix section 3.5.2. □

For using the result in Lemma 3.5.1 we need to define penalties $\text{pen}(\underline{\mathbf{X}}) \geq 1$ on candidate reconstructions $\underline{\mathbf{X}}$ of $\underline{\mathbf{X}}^*$, so that for every subset $\mathcal{X}$ of the set $\mathcal{X}'$ specified in the conditions of Theorem 4.5.1 the summability condition $\sum_{\underline{\mathbf{X}} \in \mathcal{X}} 2^{-\text{pen}(\underline{\mathbf{X}})} \leq 1$ holds. To this end, we will use the fact that for any $\mathcal{X} \subseteq \mathcal{X}'$ we always have $\sum_{\underline{\mathbf{X}} \in \mathcal{X}} 2^{-\text{pen}(\underline{\mathbf{X}})} \leq \sum_{\underline{\mathbf{X}} \in \mathcal{X}'} 2^{-\text{pen}(\underline{\mathbf{X}})}$; thus, it suffices for us to show that for the specific set $\mathcal{X}'$ described in (3.10), the penalty satisfies the Kraft-McMillan inequality:

$$\sum_{\underline{\mathbf{X}} \in \mathcal{X}'} 2^{-\text{pen}(\underline{\mathbf{X}})} \leq 1. \tag{3.20}$$

The Kraft-McMillan Inequality is automatically satisfied if we set the pen($\underline{\mathbf{X}}$) to be the code length of some uniquely decodable binary code for the elements $\underline{\mathbf{X}} \in \mathcal{X}'$ [103].

We utilize a common encoding strategy for encoding the elements of $\mathcal{A}$ and $\mathcal{B}$. We encode each entry of the matrices using $\log_2(L_{\text{lev}})$ bits in this manner the total number of bits needed to code any elements in $\mathcal{A}$ and $\mathcal{B}$ is $n_1 F \log_2(L_{\text{lev}})$ and $n_2 F \log_2(L_{\text{lev}})$ respectively. Since the elements of set $\mathcal{C}$ are sparse we follow a two step procedure: first we encode the location of the non-zero elements using $\log_2 L_{\text{loc}}$ bits where $L_{\text{loc}} = 2^{\lceil \log_2(n_3 F) \rceil}$ and then we encode the entry using $\log_2(L_{\text{lev}})$ bits. Now, we let $\mathcal{X}''$ be the set of all such $\underline{\mathbf{X}}$ with CP factors $\mathbf{A} \in \mathcal{A}$, $\mathbf{B} \in \mathcal{B}$, $\mathbf{C} \in \mathcal{C}$, and let the code for each $\underline{\mathbf{X}}$ be the concatenation of the (fixed-length) code for $\mathbf{A}$ followed by (fixed-length) code for $\mathbf{B}$ followed by the (variable-length) code for $\mathbf{C}$. It follows that we may assign penalties pen($\underline{\mathbf{X}}$) to all $\underline{\mathbf{X}} \in \mathcal{X}''$ whose lengths satisfy

$$\text{pen}(\underline{\mathbf{X}}) = (n_1 + n_2) F \log_2 L_{\text{lev}} + \|\mathbf{C}\|_0 \log_2(L_{\text{loc}} L_{\text{lev}}).$$

By construction such a code is uniquely decodable, since by the Kraft McMillan inequality we have $\sum_{X \in \mathcal{X}''} 2^{-\text{pen}(\underline{\mathbf{X}})} \leq 1$. Further, since $\mathcal{X}' \subseteq \mathcal{X}''$ (because $\mathcal{X}'$ has a constraint on entries being bounded in magnitude by X_{max}) this also satisfies the inequality $\sum_{\underline{\mathbf{X}} \in \mathcal{X}} 2^{-\text{pen}(\underline{\mathbf{X}})} \leq 1$ in (3.18) in Lemma 3.5.1 is satisfied for $\mathcal{X}'$ sa defined in statement of the Theorem 4.5.1. Now for any set $\underline{\mathbf{X}} \subseteq \mathcal{X}'$ and using coding strategy described above, the condition (3.18) in Lemma (3.5.1) is satisfied. So for randomly sub-sampled and noisy observations $\underline{\mathbf{Y}}_{\mathcal{S}}$ our estimates take the form

$$\begin{aligned}
\widehat{\underline{\mathbf{X}}}^{\xi} &= \underset{\underline{\mathbf{X}}=[\mathbf{A},\mathbf{B},\mathbf{C}] \in \mathcal{X}}{\arg\min} \left\{ -\log p_{\underline{\mathbf{X}}_{\mathcal{S}}}(\underline{\mathbf{Y}}_{\mathcal{S}}) + \xi \text{pen}(\underline{\mathbf{X}}) \right\} \\
&= \underset{\underline{\mathbf{X}}=[\mathbf{A},\mathbf{B},\mathbf{C}] \in \mathcal{X}}{\arg\min} \left\{ -\log p_{\underline{\mathbf{X}}_{\mathcal{S}}}(\underline{\mathbf{Y}}_{\mathcal{S}}) + \xi \log_2(L_{\text{loc}} L_{\text{lev}}) \|\mathbf{C}\|_0 \right\}
\end{aligned}$$

Further, when ξ satisfies (3.19), we have

$$\begin{aligned}\frac{\mathbb{E}_{\mathcal{S},\underline{\mathbf{Y}}_{\mathcal{S}}}\left[-2\log(H(p_{\hat{\underline{\mathbf{X}}}^*},p_{\underline{\mathbf{X}}^*}))\right]}{n_1n_2n_3} &\leq \frac{8Q_D\log m}{m}+\\ &\quad 3\min_{\underline{\mathbf{X}}\in\mathcal{X}}\left\{\frac{D(p_{\underline{\mathbf{X}}^*}\|p_{\underline{\mathbf{X}}})}{n_1n_2n_3}+\left(\xi+\frac{4Q_D\log 2}{3}\right)\right.\\ &\quad \left.\cdot\frac{(n_1+n_2)F\log_2 L_{\text{lev}}+\|\mathbf{C}\|_0\log_2(L_{\text{loc}}L_{\text{lev}})}{m}\right\}\\ &\leq \frac{8Q_D\log m}{m}+\\ &\quad 3\min_{\underline{\mathbf{X}}\in\mathcal{X}}\left\{\frac{D(p_{\underline{\mathbf{X}}^*}\|p_{\underline{\mathbf{X}}})}{n_1n_2n_3}+\left(\xi+\frac{4Q_D\log 2}{3}\right)\right.\\ &\qquad \left.\cdot\log_2(L_{\text{loc}}L_{\text{lev}})\frac{(n_1+n_2)F+\|\mathbf{C}\|_0}{m}\right\}.\end{aligned}$$

Finally, we let $\lambda=\xi\cdot\log_2(L_{\text{loc}}L_{\text{lev}})$ and using the relation that

$$\log_2 L_{\text{loc}}L_{\text{lev}}\leq 2\cdot(\beta+2)\cdot\log(n_{\max}) \tag{3.21}$$

which follows by our selection of L_{lev} and L_{loc} and the fact that $F,n_3\leq n_{\max}$ and $n_{\max}\geq 4$. Using the condition (3.21) and (3.19) in Lemma 3.5.1 it follows that for

$$\lambda\geq 4(\beta+2)\left(1+\frac{2Q_D}{3}\right)\log(n_{\max})$$

the estimate

$$\hat{\underline{\mathbf{X}}}^{\lambda}=\underset{\underline{\mathbf{X}}=[\mathbf{A},\mathbf{B},\mathbf{C}]\in\mathcal{X}}{\arg\min}\left(-\log p_{\underline{\mathbf{X}}_{\mathcal{S}}}(Y_{\mathcal{S}})+\lambda\|\mathbf{C}\|_0\right), \tag{3.22}$$

satisfies the bound (3.12) in Theorem 4.5.1.

3.5.2 Proof of lemma 3.5.1

The main requirement for the proof of this lemma is to show that our random Bernoulli measurement model is "good" in the sense that it will allow us to apply some known concentration results. Let Q_D be an upper bound on the KL-divergence of $p_{\mathbf{X}_{i,j,k}}$ from $p_{\underline{\mathbf{X}}^*_{i,j,k}}$ over all elements $\mathbf{X}\in\mathcal{X}$:

$$Q_D\geq\max_{\underline{\mathbf{X}}\in\mathcal{X}}\max_{i,j,k}D(p_{X^*_{i,j,k}}\|p_{X_{i,j,k}}).$$

Similarly, let Q_A be an upper bound on negative two times the log of the Hellinger affinities between the same:

$$Q_A \geq \max_{\underline{\mathbf{X}} \in \mathcal{X}} \max_{i,j,k} -2 \log \left(H(p_{X^*_{i,j,k}} \| p_{X_{i,j,k}}) \right).$$

Let $m \leq n_1 n_2 n_3$ be the expected total number of measurements and $\gamma = m/(n_1 n_2 n_3)$ to be the ratio of measured entries to total entries. Given any $\delta \in (0,1)$ define the "good" set $\mathcal{G}_{\gamma,\delta}$ as the subset of all possible sampling sets that satisfy a desired property:

$$\mathcal{G}_{\gamma,\delta} := \Bigg\{ \S \subseteq [n_1] \times [n_2] \times [n_3] : \\ \left(\bigcap_{\underline{\mathbf{X}} \in \mathcal{X}} D(p_{\underline{\mathbf{X}}^*_{\mathcal{S}}} \| p_{\underline{\mathbf{X}}_{\mathcal{S}}}) \leq \frac{3\gamma D(p_{\underline{\mathbf{X}}^*} \| p_{\underline{\mathbf{X}}})}{2} + \frac{4 Q_D [\log(1/\delta) + \log 2^{\mathrm{pen}(\underline{\mathbf{X}})}]}{3} \right) \\ \cap \left(\bigcap_{\underline{\mathbf{X}} \in \mathcal{X}} -2 \log H(p_{\underline{\mathbf{X}}^*_{\mathcal{S}}}, p_{\underline{\mathbf{X}}_{\mathcal{S}}}) \geq \frac{-2\gamma \log H(p_{\underline{\mathbf{X}}^*}, p_{\underline{\mathbf{X}}})}{2} - \frac{4 Q_A [\log(1/\delta) + \log 2^{\mathrm{pen}(\underline{\mathbf{X}})}]}{3} \right) \Bigg\}$$

We show that Bernoulli sampling with parameter γ will be "good" with high probability in the following lemma.

Lemma 3.5.2. *Let $\mathcal{X}$ be a finite collection of countable estimates $\underline{\mathbf{X}}$ for $\underline{\mathbf{X}}^*$ with penalties* $\mathrm{pen}(\underline{\mathbf{X}})$ *satisfying the Kraft inequality* (3.18). *Then for any fixed $\gamma, \delta \in (0,1)$ let $\mathcal{S}$ be a random subset of $[n_1] \times [n_2] \times [n_3]$ be a random subset generated according the Bernoulli sampling model.Then $\mathbb{P}[\mathcal{S} \notin \mathcal{G}_{\gamma,\delta}) \leq 2\delta$.*

Proof. Note that $\mathcal{G}_{\gamma,\delta}$ is defined in terms of an intersection of two events, define them to be

$$\mathcal{E}_D = \left\{ \bigcap_{\underline{\mathbf{X}} \in \mathcal{X}} D(p_{\underline{\mathbf{X}}^*_{\mathcal{S}}} \| p_{\underline{\mathbf{X}}_{\mathcal{S}}}) \leq \frac{3\gamma D(p_{\underline{\mathbf{X}}^*} \| p_{\underline{\mathbf{X}}})}{2} + \frac{4 Q_D [\log(1/\delta) + \log 2^{\mathrm{pen}(\underline{\mathbf{X}})}]}{3} \right\}$$

and

$$\mathcal{E}_A = \left\{ \bigcap_{\underline{\mathbf{X}} \in \mathcal{X}} -2 \log H(p_{\underline{\mathbf{X}}^*_{\mathcal{S}}}, p_{\underline{\mathbf{X}}_{\mathcal{S}}}) \geq \frac{-2\gamma \log H(p_{\underline{\mathbf{X}}^*}, p_{\underline{\mathbf{X}}})}{2} - \frac{4 Q_A [\log(1/\delta) + \log 2^{\mathrm{pen}(\underline{\mathbf{X}})}]}{3} \right\}$$

We apply the union bound to find that

$$\mathbb{P}\left[\mathcal{S} \notin \mathcal{G}_{\gamma,\delta}\right] \leq \mathbb{P}\left[\mathcal{E}_D^C\right] + \mathbb{P}\left[\mathcal{E}_A^C\right],$$

and will prove the theorem by showing that each of the two probabilities on the right-hand side are less than δ, starting with $\mathbb{P}[\mathcal{E}_D^C]$.

Since the observations are conditionally independent given $\mathcal{S}$, we know that for fixed $\underline{\mathbf{X}} \in \mathcal{X}$,

$$D(p_{\underline{\mathbf{X}}^*_{\mathcal{S}}} \| p_{\underline{\mathbf{X}}_{\mathcal{S}}}) = \sum_{(i,j,k)\in\mathcal{S}} D(p_{X^*_{i,j,k}} \| p_{X_{i,j,k}}) = \sum_{i,j,k} b_{i,j,k} D(p_{X^*_{i,j,k}} \| p_{X_{i,j,k}}),$$

where $b_{i,j,k} \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(\gamma)$. We will show that random sums of this form are concentrated around its mean using the Craig-Bernstein inequality .

The version of the Craig-Bernstein inequality that we will use states: let $U_{i,j,k}$ be random variables such that we have the uniform bound $|U_{i,j,k} - \mathbb{E}[U_{i,j,k}]| \leq \beta$ for all i,j,k. Let $\tau > 0$ and ϵ be such that $0 < \epsilon\beta/3 < 1$. Then

$$\mathbb{P}\left[\sum_{i,j,k}(U_{i,j,k} - \mathbb{E}[U_{i,j,k}]) \geq \frac{\tau}{\epsilon} + \epsilon \frac{\sum_{i,j,k} \text{var}(U_{i,jk})}{2(1-\epsilon\beta/3)}\right] \leq e^{-\tau}.$$

To apply the Craig-Bernstein inequality to our problem we first fix $\underline{\mathbf{X}} \in \mathcal{X}$ and define $U_{i,j,k} = b_{i,j,k} D(p_{X^*_{i,j,k}} \| p_{X_{i,j,k}})$. Note that $U_{i,j,k} \leq Q_D \Rightarrow |U_{i,j,k} - \mathbb{E}[U_{i,j,k}]| \leq Q_D$. We also bound the variance via

$$\begin{aligned}\text{var}(U_{i,j,k}) &= \gamma(1-\gamma)\left(D(p_{X^*_{i,j,k}} \| p_{X_{i,j,k}})\right)^2 \\ &\leq \gamma\left(D(p_{X^*_{i,j,k}} \| p_{X_{i,j,k}})\right)^2.\end{aligned}$$

Then let $\epsilon = \frac{3}{4Q_D}$ and $\beta = Q_D$ in (3.5.2) to get that

$$\mathbb{P}\left[\sum_{i,j,k}(b_{i,j,k} - \gamma) D(p_{X^*_{i,j,k}} \| p_{X_{i,j,k}}) \geq \frac{4Q_D\tau}{3} + \frac{\sum_{i,j,k}\gamma \cdot \left(D(p_{X^*_{i,j,k}} \| p_{X_{i,j,k}})\right)^2}{2Q_D}\right] \leq e^{-\tau}.$$

Now use the fact that $D(p_{X^*_{i,j,k}} \| p_{X_{i,j,k}}) \leq Q_D$ by definition to cancel out the square term to get:

$$\mathbb{P}\left[\sum_{i,j,k}(b_{i,j,k} - \gamma) D(p_{X^*_{i,j,k}} \| p_{X_{i,j,k}}) \geq \frac{4Q_D\tau}{3} + \frac{\gamma}{2}\sum_{i,j,k} D(p_{X^*_{i,j,k}} \| p_{X_{i,j,k}})\right] \leq e^{-\tau}.$$

Finally, we define $\delta = e^{-\tau}$, and simplify to arrive at

$$\mathbb{P}\left[D(p_{\underline{\mathbf{X}}^*_{\mathcal{S}}} \| p_{\underline{\mathbf{X}}_{\mathcal{S}}}) \geq \frac{4Q_D \log(1/\delta)}{3} + \frac{3\gamma}{2} D(p_{\underline{\mathbf{X}}^*} \| p_{\underline{\mathbf{X}}})\right] \leq \delta, \tag{3.23}$$

for any δ.

To get a uniform bound over all $\underline{\mathbf{X}} \in \mathcal{X}$ define $\delta_{\underline{\mathbf{X}}} := \delta 2^{-\mathrm{pen}(\underline{\mathbf{X}})}$ and use the bound in (3.23) with $\delta_{\underline{\mathbf{X}}}$ and apply the union bound over the class $\mathcal{X}$ to find that

$$\mathbb{P}\left[\mathcal{E}_A\right] \leq \delta. \tag{3.24}$$

An similar argument (applying Craig-Bernstein and a union bound) can be applied to $\mathcal{E}_A$ to obtain

$$\mathbb{P}\left[\mathcal{E}_A\right] \leq \delta \tag{3.25}$$

This completes the proof of lemma 3.5.2. □

Given lemma 3.5.2, the rest of the proof of lemma 3.5.1 is a straightforward extension of the published proof of lemma A.1 in [2].

3.5.3 Proof of corollary 3.3.1

We first establish a general error bound, which we then specialize to the case stated in the corollary. Note that for $\underline{\mathbf{X}}^*$ as specified and any $\underline{\mathbf{X}} \in \mathcal{X}$, using the model (3.13) we have

$$D(p_{X^*_{i,j,k}} \| p_{X_{i,j,k}}) = \frac{(X^*_{i,j,k} - X_{i,j,k})^2}{2\sigma^2}$$

for any fixed $(i,j,k) \in \mathcal{S}$. It follows that $D(p_{\underline{\mathbf{X}}^*} \| p_{\underline{\mathbf{X}}}) = \|\underline{\mathbf{X}}^* - \underline{\mathbf{X}}\|_F^2 / 2\sigma^2$. Further. as the amplitudes of entries of $\underline{\mathbf{X}}^*$ and all $\underline{\mathbf{X}} \in \mathcal{X}$ upper bounded by $X_{\max}$, it is easy to see that we may choose $Q_D = 2X_{\max}^2/\sigma^2$. Also, for any $\underline{\mathbf{X}} \in \mathcal{X}$ and any fixed $(i,j,k) \in \mathcal{S}$ it is easy to show that in this case

$$-2\log H(p_{X_{i,j,k}}, p_{X^*_{i,j,k}}) = \frac{(X^*_{i,j,k} - X_{i,j,k})^2}{4\sigma^2},$$

so that $-2\log H(p_{\underline{\mathbf{X}}}, p_{\underline{\mathbf{X}}^*}) = \|\underline{\mathbf{X}}^* - \underline{\mathbf{X}}\|_F^2/4\sigma^2$. It follows that

$$\mathbb{E}_{\mathcal{S},\underline{\mathbf{Y}}_{\mathcal{S}}}\left[-2\log H(p_{\hat{\underline{\mathbf{X}}}}, p_{\underline{\mathbf{X}}^*})\right] = \frac{\mathbb{E}_{\mathcal{S},\underline{\mathbf{Y}}_{\mathcal{S}}}\left[\|\underline{\mathbf{X}}^* - \widehat{\underline{\mathbf{X}}}\|_F^2\right]}{4\sigma^2}.$$

Now for using Theorem 4.5.1, we first substitute the value of $Q_D = 2X_{\max}^2/\sigma^2$ to obtain the following condition on λ

$$\lambda \geq 4 \cdot \left(1 + \frac{4X_{\max}^2}{3\sigma^2}\right) \cdot (\beta + 2) \cdot \log(n_{\max}).$$

The above condition implies that the specific choice of λ given (3.15) is a valid choice to use if we want to invoke Theorem 4.5.1. So fixing λ as given (3.15) and using Theorem 4.5.1, the sparsity penalized ML estimate satisfies the per-element mean-square error bound

$$\frac{\mathbb{E}_{\mathcal{S},\underline{\mathbf{Y}}_{\mathcal{S}}}\left[\|\underline{\mathbf{X}}^* - \widehat{\underline{\mathbf{X}}}\|_F^2\right]}{n_1 n_2 n_3} \leq \frac{64 X_{\max}^2 \log m}{m} + \\ 6 \cdot \min_{\underline{\mathbf{X}} \in \mathcal{X}} \left\{ \frac{\|\underline{\mathbf{X}}^* - \underline{\mathbf{X}}\|_F^2}{n_1 n_2 n_3} + \left(2\sigma^2\lambda + \frac{24 X_{\max}^2 (\beta+2)\log(n_{\max})}{3}\right)\left(\frac{(n_1+n_2)F + \|\mathbf{C}\|_0}{m}\right)\right\}.$$

Notice that the above inequality is sort of an oracle type inequality because it implies that for any $\underline{\mathbf{X}} \in \mathcal{X}$ we have

$$\frac{\mathbb{E}_{\mathcal{S},\underline{\mathbf{Y}}_{\mathcal{S}}}\left[\|\underline{\mathbf{X}}^* - \widehat{\underline{\mathbf{X}}}\|_F^2\right]}{n_1 n_2 n_3} \leq \frac{64 X_{\max}^2 \log m}{m} + \\ 6 \cdot \left\{ \frac{\|\underline{\mathbf{X}}^* - \underline{\mathbf{X}}\|_F^2}{n_1 n_2 n_3} + \left(2\sigma^2\lambda + \frac{24 X_{\max}^2 (\beta+2)\log(n_{\max})}{3}\right)\left(\frac{(n_1+n_2)F + \|\mathbf{C}\|_0}{m}\right)\right\}.$$

We use this inequality for a specific candidate reconstruction of form $\underline{\mathbf{X}}_Q^* = [\mathbf{A}_Q^*, \mathbf{B}_Q^*, \mathbf{C}_Q^*]$ where the entries of $\mathbf{A}_Q^*$ are the closest discretized surrogates of the entries of $\mathbf{A}^*$, $\mathbf{B}_Q^*$ are the closest discretized surrogates of the entries of $\mathbf{B}^*$, and $\mathbf{C}_Q^*$ are the closest discretized surrogates of the non-zeros entries of $\mathbf{C}^*$ (and zero otherwise). For proceeding further we need to bound $\|\underline{\mathbf{X}}_Q^* - \underline{\mathbf{X}}^*\|_{\max}$. For this purpose we consider matricization of tensor across the third dimension as follows

$$\|\underline{\mathbf{X}}_Q^* - \underline{\mathbf{X}}^*\|_{\max} = \left\|\left(\mathbf{B}_Q^* \odot \mathbf{A}_Q^*\right)(\mathbf{C}_Q^*)^T - (\mathbf{B}^* \odot \mathbf{A}^*)(\mathbf{C}^*)^T\right\|_{\max}$$

Next we write $\mathbf{A}_Q^* = \mathbf{A}^* + \Delta_A$, $\mathbf{B}_Q^* = \mathbf{B}^* + \Delta_B$ and $\mathbf{C}_Q^* = \mathbf{C}^* + \Delta_C$ with straight forward matrix multiplication we can obtain that

$$\begin{aligned}\left(\mathbf{B}_Q^* \odot \mathbf{A}_Q^*\right)(\mathbf{C}_Q^*)^T =& (\mathbf{B}^* \odot \mathbf{A}^*)(\mathbf{C}^*)^T + (\Delta_A \odot \mathbf{B}^* + \mathbf{A}^* \odot \Delta_B + \Delta_A \odot \Delta_B)(\mathbf{C}^*)^T \\ &+ (\mathbf{A}^* \odot \mathbf{B}^* + \Delta_A \odot \mathbf{B}^* + \mathbf{A}^* \odot \Delta_B + \Delta_A \odot \Delta_B)\Delta_C^T\end{aligned}$$

Using this identity it follows

$$\begin{aligned}\|\underline{\mathbf{X}}_Q^* - \underline{\mathbf{X}}^*\|_{\max} = \|&(\Delta_A \odot \mathbf{B}^* + \mathbf{A}^* \odot \Delta_B + \Delta_A \odot \Delta_B)(\mathbf{C}^*)^T + \\ &(\mathbf{A}^* \odot \mathbf{B}^* + \Delta_A \odot \mathbf{B}^* + \mathbf{A}^* \odot \Delta_B + \Delta_A \odot \Delta_B)\Delta_C^T\|_{\max}\end{aligned}$$

Now using the facts that $\|\mathbf{A} \odot \mathbf{B}\|_{\max} = \|\mathbf{A}\|_{\max}\|\mathbf{B}\|_{\max}$, $\|\mathbf{AB}\|_{\max} \leq F\|\mathbf{A}\|\|\mathbf{B}\|_{\max}$ and triangle inequality for the $\|\cdot\|_{\max}$ norm it is easy to show that

$$\|\underline{\mathbf{X}}_Q^* - \underline{\mathbf{X}}^*\|_{\max} \leq F[(\|\Delta_A\|_{\max} + \|\mathbf{A}\|_{\max})(\|\Delta_B\|_{\max}+\|\mathbf{B}\|_{\max})(\|\Delta_C\|_{\max} + \|\mathbf{C}\|_{\max}) \\ -\|\mathbf{A}\|_{\max}\|\mathbf{B}\|_{\max}\|\mathbf{C}\|_{\max}]$$

Further, using the fact that $\|\Delta_A\|_{\max} \leq \frac{A_{\max}}{L_{\text{lev}}-1}$, $\|\Delta_B\|_{\max} \leq \frac{B_{\max}}{L_{\text{lev}}-1}$, and $\|\Delta_C\|_{\max} \leq \frac{C_{\max}}{L_{\text{lev}}-1}$, we have

$$\begin{aligned}
&\|\underline{\mathbf{X}}_Q^* - \underline{\mathbf{X}}^*\|_{\max} \\
&\leq F\left[\left(\frac{A_{\max}}{L_{\text{lev}}-1} + A_{\max}\right)\left(\frac{B_{\max}}{L_{\text{lev}}-1} + \|B\|_{\max}\right)\left(\frac{C_{\max}}{L_{\text{lev}}-1} + C_{\max}\right) - A_{\max}B_{\max}C_{\max}\right] \\
&\leq FA_{\max}B_{\max}C_{\max}\left[\left(1 + \frac{1}{L_{\text{lev}}-1}\right)^3 - 1\right] \\
&\leq \frac{FA_{\max}B_{\max}C_{\max}}{L_{\text{lev}}-1}\left[3 + \frac{3}{L_{\text{lev}}-1} + \frac{1}{(L_{\text{lev}}-1)^2}\right] \\
&\leq \frac{7FA_{\max}B_{\max}C_{\max}}{L_{\text{lev}}-1},
\end{aligned}$$

where in the second last step we have used $L_{\text{lev}} \geq 2$. Now, it is straight-forward to show that our choice of β in (3.14) implies $L_{\text{lev}} \geq 14FA_{\max}B_{\max}C_{\max}/X_{\max}+1$, so each entry of $\|\underline{\mathbf{X}}_Q^* - \underline{\mathbf{X}}^*\|_{\max} \leq X_{\max}/2$. This further implies that for the candidate estimate $\underline{\mathbf{X}}_Q^*$ we have $\|\underline{\mathbf{X}}_Q^*\|_{\max} \leq X_{\max}$, i.e., $\underline{\mathbf{X}}_Q^* \in \mathcal{X}$. Moreover, we

$$\frac{\|\underline{\mathbf{X}}^* - \underline{\mathbf{X}}_Q^*\|_F^2}{n_1n_2n_3} \leq \left(\frac{7FA_{\max}B_{\max}C_{\max}}{L_{\text{lev}}-1}\right)^2 \leq \frac{X_{\max}^2}{m}, \tag{3.26}$$

where the last inequality follows from the fact that our specific choice of β in (3.14) also implies $L_{\text{lev}} \geq 7F\sqrt{m}A_{\max}B_{\max}C_{\max}/X_{\max}$.

Finally, we evaluate the oracle inequality for (4.5) for $\underline{\mathbf{X}}_Q^*$ and using the fact that $\|C_Q^*\|_0 = \|C^*\|_0$ and using the value of λ specified in the corollary we have

$$\frac{\mathbb{E}_{\mathcal{S},\underline{\mathbf{Y}}_{\mathcal{S}}}\left[\|\underline{\mathbf{X}}^* - \widehat{\underline{\mathbf{X}}}\|_F^2\right]}{n_1n_2n_3} \leq \\ \frac{70X_{\max}^2\log m}{m} + 24(\sigma^2 + 2X_{\max}^2)(\beta+2)\log(n_{\max})\left(\frac{(n_1+n_2)F + \|\mathbf{C}^*\|_0}{m}\right).$$

Chapter 4

Matrix completion from noisy and quantized observations

As discussed in Chapter 3 matrix completion problem arises in variety of signal processing and machine learning applications. Noisy and quantized observations usually arise in systems where data is collected via low-cost resource constrained devices. Sometimes the quantization is inherent to the observation setup. For example, consider the *collaborative filtering* task which can be posed as matrix completion of matrix of size $n_1 \times n_2$ whose $(i, j)^{th}$ entry which contain rating of user i^{th} for j^{th} item. In this case typically ratings are quantized to finite symbols like number of stars or thumbs up/down.

In existing literature, the problem of matrix completion with quantized data has been primarily explored for low rank matrices. One of the initial works [98] focused on matrix completion with 1-bit data for low rank matrices with bounded entries using nuclear-norm constrained maximum likelihood type approach. In [104] authors the considered the same problem using max-norm constrained minimization approach. These works were extended to multi-bit scenarios in [79] where authors proposed numerical algorithms and error bounds were investigated in [105–107]. Surprisingly, to the best of our knowledge with the notable exception of [2, 108–110] matrix completion with quantization has not been studied for structures other than sets of low rank matrices.

The investigation of quantized matrix completion for general structure is a problem of great practical significance. In light of limited existing works, on general quantized matrix completion it requires a systematic investigation. We assume that the true matrix $\mathbf{X}^*$ lies in some set $\mathcal{X} \subset \left\{\mathbf{X} \mid \mathbf{X} \in \mathbb{R}^{n_1 \times n_2}, \|\mathbf{X}\|_\infty \leq x_{max}\right\}$, here $\|\mathbf{X}\|_\infty$ is the maximum absolute entry of the matrix. The set $\mathcal{X}$ may vary depending on the specific application at hand. We assume that the noise corrupted entries of $\mathbf{X}^*$ observed at the subset of locations and are quantized to one of K-symbols via a quantizer. We obtain estimates of true matrix $\mathbf{X}^*$ from these noise corrupted quantized by solving a constrained maximum likelihood problem. Our main contribution here comes in the form two general probabilistic error guarantees for the constrained maximum likelihood estimates obtained via covering number based approach, and a more involved chaining principle based approach. We demonstrate the utility of these bounds by instantiating them for the set of low rank matrices as well as matrices following the sparse factor model. For the completion of matrices following the sparse factor model we propose ab alternating direction method of multiplier-type algorithm for approximately solving the constrained maximum likelihood problem and provide empirical evidence for the theoretical bounds.

4.1 Preliminaries and notations

For a positive integer n, we let $[n] = \{1, 2, \ldots, n\}$. We use the asterisk superscript to denote that the corresponding parameter is the "true" model in our estimation task, rather than to denote complex conjugation. We denote scalars with lower case letters (e.g., x) and vectors with bold face letters (e.g., $\mathbf{x}$). The matrices are denoted by upper case bold face letters (e.g., $\mathbf{X}$) and $(i,j)^{th}$ entry of the matrix $\mathbf{X}$ is denote by $X_{i,j}$. The norm $\|\mathbf{X}\|_\infty$ of matrix $\mathbf{X}$ is equal to the maximum absolute value of its entries. The probability that a discrete variable Z takes the value k from possible K different values is denoted by $p(Z = k)$, the probability mass function (pmf) is denoted by the shorthand notation $p(Z)$, and KL-divergence between the two pmfs $p(Z)$ and $q(Z)$ is defined as

$$\mathrm{D}\left(p(Z) || q(Z)\right) = \sum_{i=1}^{K} p(i) \log\left(\frac{p(i)}{q(i)}\right).$$

4.2 Observation model

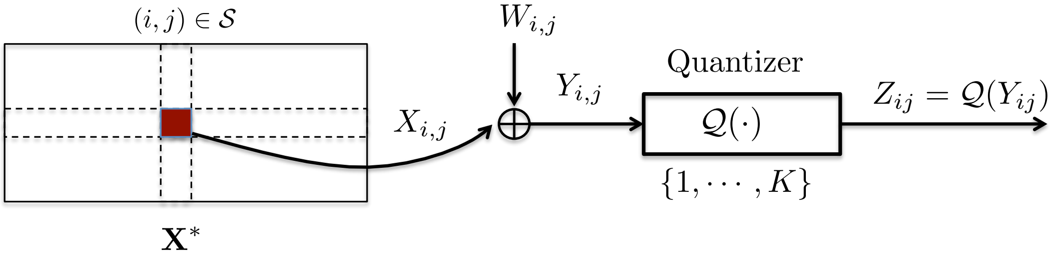


Figure 4.1: Observation model under quantized setting

Suppose we observe the entires of matrix $\mathbf{X}^*$ at a random subset of entries $\mathcal{S} \subset [n_1] \times [n_2]$ chosen such that each entry is observed independently and identically with probability $\gamma \in (0,1)$. The noise corrupted entries are first obtained as follows

$$Y_{i,j} = X^*_{i,j} + W_{i,j},\ (i,j,) \in \mathcal{S}$$

where the $\{W_{i,j}\}_{(i,j)\in\mathcal{S}}$ denotes the noise. Subsequently, these noisy corrupted entries are passed through the quantizer $\mathcal{Q}(\cdot)$ that quantizes them to K symbols to obtain the final quantized obervations as follows

$$Z_{i,j} = \mathcal{Q}(Y_{i,j}) = \begin{cases} 1 & -\infty < Y_{i,j} <= \tau_1 \\ k & \tau_{k-1} < Y_{i,j} <= \tau_k,\ k = 2, \cdots, K-1 \\ K & \tau_{K-1} < Y_{i,j} < +\infty \end{cases}, \quad (i,j) \in \mathcal{S} \tag{4.1}$$

where $\{\tau_1, \cdots, \tau_{K-1}\}$ are the given thresholds of the quantizer. The final observation model is shown in Figure 4.1. We assume that all noise $W_{i,j}$ are i.i.d. zero-mean random variables distributed with common probability density function (pdf) $f_W(w)$ and cumulative distribution function (cdf) $F_W(w)$.

4.3 Estimation procedure

Our estimation approach will be based on a variant of the well-known maximum likelihood approach. We first obtain the joint probability of the observation. The independence of noises $\{W_{i,j}\}_{(i,j)\in\mathcal{S}}$ implies that the observations $\{Z_{i,j}\}_{(i,j)\in\mathcal{S}}$ are also independent. Each quantized observation $Z_{i,j}$ is a discrete random variables whose pmf is parametrized by the entry $X^*_{i,j}$ of the true matrix as

$$p(k; X^*_{i,j}) = Pr(Z_{i,j} = k; X^*_{i,j}) = \int_{\tau_{k-1}-X^*_{i,j}}^{\tau_k - X^*_{i,j}} f_W(w)dw, \quad k = 1, \cdots, K \tag{4.2}$$

where $\tau_0 = -\infty$ and $\tau_K = +\infty$. Denoting the observed entries of $\{Z_{i,j}\}_{(i,j)\in\mathcal{S}}$ by $\mathbf{Z}_{\mathcal{S}}$, the joint pmf for $\mathbf{Z}_{\mathcal{S}}$ is given by

$$p(\mathbf{Z}_{\mathcal{S}}; \mathbf{X}^*_{\mathcal{S}}) = \prod_{(i,j)\in\mathcal{S}} p(Z_{i,j}; X^*_{i,j}),$$

where

$$p(Z_{i,j}; X^*_{i,j}) = \prod_{k=1}^{K} Pr(Z_{i,j} = k; X^*_{i,j})^{\mathbf{1}(Z_{i,j}=k)},$$

with $\mathbf{1}(Z_{i,j} = k) = 1$ if $Z_{i,j} = k$ otherwise $\mathbf{1}(Z_{i,j} = k) = 0$. Using the joint pmf our estimation approach involves solving the following constrained maximum log likelihood problem to obtain the estimate $\hat{\mathbf{X}}$ of true matrix $\mathbf{X}^*$ as follows

$$\widehat{\mathbf{X}} = \arg\max_{\mathbf{X}\in\mathcal{X}} \ \log p(\mathbf{Z}_{\mathcal{S}}; \mathbf{X}_{\mathcal{S}}). \tag{4.3}$$

An alternate expression for the objective function in the above constrained maximum log likelihood problem in terms of sampling set $\mathcal{S}$ will be used in subsequent sections. As mentioned earlier, we assume the random sampling model in which each entry is observed independently and identically with probability $\gamma \in (0, 1)$. The sampling of $(i, j)^{th}$ entry can then be indicated by a binary random variable $b_{ij} \sim Bernoulli(\gamma)$. With this, the sampling set can be written in be terms of $b_{i,j}$'s as $\mathcal{S} = \{(i, j)|b_{i,j} = 1\}$. In terms of $b_{i,j}$ and observations $\mathbf{Z}_{\mathcal{S}}$ the objective of the constrained maximum likelihood problem in 4.3 can be written as

$$\log p(\mathbf{Z}_{\mathcal{S}}; \mathbf{X}_{\mathcal{S}}) = \sum_{i,j} b_{ij} \log p(Z_{ij}; X_{ij})$$

The numerical algorithm to solve the constrained maximum log likelihood problem in (4.3) depends on the constraint set $\mathcal{X}$. Depending on the specific set $\mathcal{X}$ this problem may be convex or non-convex. We discuss numerical algorithms to solve the problem (4.3) in section 4.8.

4.4 Towards error guarantees

The estimate $\hat{\mathbf{X}}$ obtained via constrained maximum likelihood approach in (4.3) is a random variable which could potentially lie anywhere in the set $\mathcal{X}$ depending on the random realization of the sampling set $\mathcal{S}$, noise, and $\mathbf{X}^*$. Therefore, an important theoretical question is about the quality of estimate $\hat{\mathbf{X}}$. The random variables are known to concentrate around their expected value which means that with high probability the random variables assumes values near their expected values. We use this fundamental nature of random variables to obtain probabilistic error guarantees for $\hat{\mathbf{X}}$ in (4.3).

For obtaining the error guarantees relative to $\mathbf{X}^*$ we focus on the following random variable indexed by a fixed $\mathbf{X} \in \mathcal{X}$

$$U_{\mathbf{X}} = \log \frac{p(\mathbf{Z}_{\mathcal{S}}; \mathbf{X}^*_{\mathcal{S}})}{p(\mathbf{Z}_{\mathcal{S}}; \mathbf{X}_{\mathcal{S}})} = \sum_{i,j} b_{ij} \log \frac{p(Z_{ij}; X^*_{ij})}{p(Z_{ij}; X_{ij})}. \tag{4.4}$$

It is easy to see that the expected value $U_{\mathbf{X}}$ is given by

$$\mathbb{E} \log \frac{p(\mathbf{Z}_{\mathcal{S}}; \mathbf{X}_{\mathcal{S}})}{p(\mathbf{Z}_{\mathcal{S}}; \mathbf{X}^*_{\mathcal{S}})} = \gamma \sum_{i,j} \mathrm{D}((p(Z_{ij}; X^*_{ij})||p(Z_{ij}; X_{ij}))) = \gamma D(p(\mathbf{Z}; \mathbf{X}^*)||p(\mathbf{Z}; \mathbf{X})),$$

where the expectation is with respect to b_{ij} and Z_{ij}, and we have used the shorthand notation $D(p(\mathbf{Z}; \mathbf{X}^*)||p(\mathbf{Z}; \mathbf{X})) = \sum_{i,j} \mathrm{D}(p(Z_{ij}; X^*_{ij})||p(Z_{ij}; X_{ij}))$. The quantity $D(p(\mathbf{Z}; \mathbf{X}^*)||p(\mathbf{Z}; \mathbf{X}))$ is a measure of closeness between the pmfs parametrized by $\mathbf{X}$ and $\mathbf{X}^*$, and if $\mathbf{X}^* = \mathbf{X}$ then $D(p(\mathbf{Z}; \mathbf{X}^*)||p(\mathbf{Z}; \mathbf{X})) = 0$. We follow an approach where we first obtain probabilistic concentration bounds that hold uniformly over the $\mathcal{X}$. These bounds bound the deviation of $U_{\mathbf{X}}$ from its expected value for all $\mathbf{X} \in \mathcal{X}$. Next we instantiate these bounds for constrained maximum likelihood estimate $\hat{\mathbf{X}}$ to obtain the final probabilistic error guarantees. There are various approaches to obtain the probabilistic concentration bounds that hold uniformly over the $\mathcal{X}$. In next two sections we follow two approaches to obtain two versions of error guarantees.

4.5 Error guarantees based on covering number approach

The random variable $U_{\mathbf{X}}$ in (4.4) for various $\mathbf{X} \in \mathcal{X}$ is a random process indexed by elements of the set $\mathcal{X}$. In the covering number based approach we first obtain concentration bound that hold for a fixed $\mathbf{X} \in \mathcal{X}$. We then consider an ϵ-cover $\mathcal{X}_\epsilon$ of the set $\mathcal{X}$, by which we mean that it is subset of $\mathcal{X}$ with minimum cardinality for which $\mathcal{X} = \bigcup_{\mathcal{X}_i \in \mathcal{X}_\epsilon} B_\epsilon(\mathbf{X}_i)$ where $B_\epsilon(\mathbf{X}_i)$ is ℓ_∞ norm ball centered at $\mathbf{X}_i \in \mathcal{X}$ with radius ϵ. We then instantiate concentration bound over the centers of the ϵ-cover and extended the argument over the entire set $\mathcal{X}$ using the Lipschitz continuity of $U_{\mathbf{X}}$ over $\mathcal{X}$.

As a first step towards obtaining performance error guarantees we establish the following concentration bound $U_{\mathbf{X}}$ for a fixed $\mathbf{X} \in \mathcal{X}$.

Lemma 4.5.1. *Suppose that $p(Z_{i,j}; X_{i,j}) \geq \delta_0$, for $\forall (i,j) \in [n_1] \times [n_2]$ and $\forall \mathbf{X} \in \mathcal{X}$ (where $0 < \delta_0 < \frac{1}{2}$). The random variable $U_{\mathbf{X}} = \sum_{i,j} b_{ij} \log \frac{p(Z_{i,j}; X^*_{ij})}{p(Z_{i,j}; X_{ij})}$ for a fixed $\mathbf{X} \in \mathcal{X}$ satisfies*

$$\mathit{Prob}\left(\frac{1}{2}\mathbb{E}(U_{\mathbf{X}}) - U_{\mathbf{X}} \geq \frac{\tau}{c}\right) \leq e^{-\tau},$$

where $c = \frac{\delta_0}{4(1-\delta_0)}$

Proof. The proof is outlined in section 4.11.2. □

Using the above concentration inequality for the covering set of $\mathcal{X}$ and using a simple union bound in addition to the Lipschtiz continuity of $U_{\mathbf{X}}$ we obtain the following probabilistic bound that holds uniformly over entire $\mathcal{X}$.

Lemma 4.5.2. *For any $\mathcal{X} \subset \left\{\mathbf{X} \;\middle|\; \mathbf{X} \in \mathbb{R}^{n_1 \times n_2}, \|\mathbf{X}\|_\infty \leq x_{max}\right\}$, let $\mathcal{X}_\epsilon = \{\mathbf{X}_1, \cdots, \mathbf{X}_N\}$ be the subset of $\mathcal{X}$ of minimum cardinality $N(\epsilon, \|\cdot\|_\infty, \mathcal{X})$ such that $\mathcal{X} = \bigcup_{\mathbf{X}_i \in \mathcal{X}_\epsilon} B_\epsilon(\mathbf{X}_i)$ where $B_\epsilon(\mathbf{X}_i)$ is ℓ_∞ norm ball centered at $\mathbf{X}_i \in \mathcal{X}_\epsilon$ with radius ϵ. Assuming $p(Z_{i,j}; X_{i,j}) \geq \delta_0$ for $\forall (i,j)$ and $\forall \mathbf{X} \in \mathcal{X}$, where $0 < \delta_0 < \frac{1}{2}$. Then the random variables $U_{\mathbf{X}} = \sum_{i,j} b_{ij} \log \frac{p(Z_{i,j}; X^*_{ij})}{p(Z_{i,j}; X_{ij})}$ satisfy the following with probability at least $1 - \alpha$*

$$\frac{1}{2}\mathbb{E}(U_{\mathbf{X}}) - U_{\mathbf{X}} \leq \left(|\mathcal{S}| + \frac{\gamma n_1 n_2}{2}\right) L_g \epsilon + \frac{\log \frac{N(\epsilon, \|\cdot\|_\infty, \mathcal{X})}{\alpha}}{c}$$

for all $\mathbf{X} \in \mathcal{X}$. Here $c = \frac{\delta_0}{4(1-\delta_0)}$ and L_g is a constant defined as follows

$$L_g = \max_k \sup_{|t| \leq x_{max}} \left| \frac{f_W(\tau_k - t) - f_W(\tau_{k-1} - t)}{\int_{\tau_{k-1}-t}^{\tau_k - t} f_W(w) dw} \right|.$$

Proof. The proof is outlined in section 4.11.3 □

Using lemma 4.5.2 we obtain the first result for the performance error guarantee on the constrained maximum likelihood estimator $\hat{\mathbf{X}}$ in (4.3) as follows

Theorem 4.5.1. *Suppose $\mathcal{S}$ is chosen such that each entry is sampled independently with probability γ such that it satisfies $\gamma n_1 n_2 \geq 12\log(\frac{2}{\alpha})$. Given the observation of $\mathbf{X}^* \in \mathcal{X} \subset \left\{\mathbf{X} \mid \mathbf{X} \in \mathbb{R}^{n_1 \times n_2}, \|\mathbf{X}\|_\infty \leq x_{max}\right\}$ taken as per (4.1) at subset of locations denoted by $\mathcal{S} \subset [n_1] \times [n_2]$. Assuming that the discrete pmf satisfies $p(Z_{i,j}; X_{i,j}) \geq \delta_0$ for $\forall (i,j)$ and $\forall$ $\mathbf{X} \in \mathcal{X}$ (where $0 < \delta_0 < \frac{1}{2}$), then with probability at least $1 - 2\alpha$ the constrained maximum likelihood estimate $\hat{\mathbf{X}}$ in (4.3) satisfies*

$$\frac{D(p(\mathbf{Z}; \mathbf{X}^*)||p(\mathbf{Z}; \hat{\mathbf{X}}))}{n_1 n_2} \leq 4L_g\epsilon + 2\frac{\log\left(\frac{2N(\epsilon, \|\cdot\|_\infty, \mathcal{X})}{\alpha}\right)}{c\gamma n_1 n_2} \tag{4.5}$$

where $N(\epsilon, \|\cdot\|_\infty, \mathcal{X})$ is the covering number of set $\mathcal{X}$ in $\|\cdot\|_\infty$ norm with accuracy ϵ,$c = \frac{\delta_0}{4(1-\delta_0)}$, and L_g defined as follows

$$L_g = max_k \sup_{|t| \leq x_{max}} \left| \frac{f_W(\tau_k - t) - f_W(\tau_{k-1} - t)}{\int_{\tau_{k-1}-t}^{\tau_k - t} f_W(w)dw} \right|. \tag{4.6}$$

Proof. Since $\hat{\mathbf{X}}$ lies in $\mathcal{X}$, using the lemma 4.5.2 we have with probability $\geq 1 - \alpha$

$$\frac{1}{2}\mathbb{E}\log\frac{p(\mathbf{Z}_\mathcal{S}; \mathbf{X}^*_\mathcal{S})}{p(\mathbf{Z}_\mathcal{S}; \hat{\mathbf{X}}_\mathcal{S})} \leq \log\frac{p(\mathbf{Z}_\mathcal{S}; \mathbf{X}^*_\mathcal{S})}{p(\mathbf{Z}_\mathcal{S}; \hat{\mathbf{X}}_\mathcal{S})} + \left(|\mathcal{S}| + \frac{\gamma n_1 n_2}{2}\right) L_g\epsilon + \frac{\log\left(\frac{2N(\epsilon, \|\cdot\|_\infty, \mathcal{X})}{\alpha}\right)}{c},$$

where we have substituted $U_{\hat{\mathbf{X}}} = \log\frac{p(\mathbf{Z}_\mathcal{S}; \mathbf{X}^*_\mathcal{S})}{p(\mathbf{Z}_\mathcal{S}; \hat{\mathbf{X}}_\mathcal{S})}$. Since $\hat{\mathbf{X}}$ maximizes the constrained maximum likelihood and $\mathbf{X}^* \in \mathcal{X}$ we have that $\log\frac{p(\mathbf{Z}_\mathcal{S}; \mathbf{X}^*_\mathcal{S})}{p(\mathbf{Z}_\mathcal{S}; \hat{\mathbf{X}}_\mathcal{S})} \leq 0$ and using the fact $\mathbb{E}\log\frac{p(\mathbf{Z}_\mathcal{S}; \mathbf{X}^*_\mathcal{S})}{p(\mathbf{Z}_\mathcal{S}; \hat{\mathbf{X}}_\mathcal{S})} = \gamma D(p(\mathbf{Z}; \mathbf{X}^*)||p(\mathbf{Z}; \hat{\mathbf{X}}))$ the above inequality reduces to

$$\frac{1}{2}\gamma D(p(\mathbf{Z}; \mathbf{X}^*)||p(\mathbf{Z}; \hat{\mathbf{X}})) \leq \left(|\mathcal{S}| + \frac{\gamma n_1 n_2}{2}\right) L_g\epsilon + \frac{\log\left(\frac{2N(\epsilon, \|\cdot\|_\infty, \mathcal{X})}{\alpha}\right)}{c}$$

Next, using lemma 4.11.5 we have that if $\gamma n_1 n_2 \geq 12\log(\frac{2}{\alpha})$ then

$$\mathrm{P}\left(\frac{1}{2}\gamma n_1 n_2 \leq |\mathcal{S}| \leq \frac{3}{2}\gamma n_1 n_2\right) \geq 1 - \alpha.$$

Assuming $\gamma n_1 n_2 \geq 12\log(\frac{2}{\alpha})$, and using union bound we have with probability at least $\geq 1 - 2\alpha$

$$\frac{D(p(\mathbf{Z}; \mathbf{X}^*)||p(\mathbf{Z}; \hat{\mathbf{X}}))}{n_1 n_2} \leq 4L_g\epsilon + 2\frac{\log\left(\frac{2N(\epsilon, \|\cdot\|_\infty, \mathcal{X})}{\alpha}\right)}{c\gamma n_1 n_2}$$

□

The main advantage of the above theorem is its generality and it can be instantiated for different choices of structured matrix sets $\mathcal{X}$ by substituting the expression for covering number $N(\epsilon, \|\cdot\|_\infty, \mathcal{X})$ for judiciously chosen ϵ. Next we give results obtained by using this strategy for few specific structured matrix sets.

4.5.1 Set of low rank matrix with bounded entries

Here we instantiate theorem 4.5.1 for the set of low rank matrices with bounded entries defined as

$$\mathcal{X} = \{\mathbf{DA}\big|\mathbf{D} \in \mathbb{R}^{n_1 \times r}, \|\mathbf{D}\|_\infty \leq d_{max}, \mathbf{A} \in \mathbb{R}^{r \times n_2}, \|\mathbf{A}\|_\infty \leq a_{max}\}, \tag{4.7}$$

where $r < \min\{n_1, n_2\}$. Note that the matrices in set $\mathcal{X}$ are not explicitly bounded in magnitude. In the above definition set of $\mathcal{X}$ we have $x_{max} = rd_{max}a_{max}$.

Corollary 4.5.1. *Under the assumptions of theorem 4.5.1. For the set of low rank matrices $\mathcal{X}$ defined in* (4.7) *given the noise corrupted quantized observation of $\mathbf{X}^* \in \mathcal{X}$ taken as per* (4.1) *then with probability at least $1-2\alpha$ the constrained maximum likelihood estimate $\hat{\mathbf{X}}$ in* (4.3) *satisfies*

$$\frac{D(p(\mathbf{Z}; \mathbf{X}^*)||p(\mathbf{Z}; \hat{\mathbf{X}}))}{n_1 n_2} \leq \frac{4L_g c^{-1} + 2(n_1 + n_2)r \log\left(12 a_{max} d_{max} r c \alpha^{-1} \gamma n_1 n_2\right) c^{-1}}{\gamma n_1 n_2},$$

where c and L_g as defined in theorem 4.5.1.

Proof. The proof is straightforward by choosing $\epsilon = \frac{1}{c\gamma n_1 n_2}$ and substituting the expression for resulting covering number in theorem 4.5.1 from lemma 4.11.6 and using the simple fact that $2\log\left(\frac{2}{\alpha}\right) \leq 2(n_1 + n_2)r \log\left(\frac{2}{\alpha}\right)$. □

The bound in the above corollary reveals that the upper bound on the error measured in terms of per-entry KL-divergence decays at the rate of γ^{-1} and is proportional to $(n_1 + n_2)r$ which may be interpreted as the degree of freedom for low rank matrix set in (4.7). Our result here can be compared to the error bounds for 1-bit matrix completion via constrained maximum likelihood estimation reported in [98] where constraint set consisted of matrices with bounded nuclear norm and bounded amplitude. In [98] the error measured in terms of per entry squared error between the true matrix $\mathbf{X}^*$ and their constrained maximum likelihood estimate is shown to be proportional to

$\sqrt{\frac{(n_1+n_2)r}{\gamma n_1 n_2}}$. This implies that bounds in [98] decay with the sampling rate as $\gamma^{0.5}$. Similarly, in [104] a decay rate of $\gamma^{0.5}$ with max-norm constrained maximum likelihood estimator. The error bound for constrained maximum likelihood estimation with exact low rank constraints reported in [107] has the rate of γ^{-1} for multi-bit quantization scenario as well. In contrast to these works, the bound in the above corollary reveals a faster rate of decay γ^{-1} although we measure the error in terms of per-entry KL-divergence and the low rank constraint is explicitly enforced. In section 4.7 we discuss a possible way to convert the bound in corollary 4.5.1 in terms of per-entry squared error.

4.5.2 Sparse factor model with bounded entries

Here we instantiate theorem 4.5.1 for the set of matrices following sparse factor model with bounded entries defined as

$$\mathcal{X} = \{\mathbf{DA}\big|\mathbf{D} \in \mathbb{R}^{n_1\times r}, \|\mathbf{D}\|_\infty \leq d_{max}, \mathbf{A} \in \mathbb{R}^{r\times n_2}, \|\mathbf{A}\|_\infty \leq a_{max}, \|\mathbf{A}\|_0 \leq l\} \tag{4.8}$$

Corollary 4.5.2. *Under the assumptions of theorem 4.5.1. For the matrix set of matrices following sparse factor model with bounded entries defined in* (4.8) *given the noise corrupted quantized observation of* $\mathbf{X}^* \in \mathcal{X}$ *taken as per* (4.1) *then with probability atleast* $1 - 2\alpha$ *the constrained maximum likelihood estimate* $\hat{\mathbf{X}}$ *in* (4.3) *satisfies*

$$\frac{D(p(\mathbf{Z};\mathbf{X}^*)||p(\mathbf{Z};\hat{\mathbf{X}}))}{n_1 n_2} \leq \frac{4L_g c^{-1} + 2(n_1 r + l)\log\left(12 a_{max} d_{max} \alpha^{-1} r^2 c\gamma n_1 n_2^2\right) c^{-1}}{\gamma n_1 n_2}$$

where c *and* L_g *as defined in theorem 4.5.1.*

Proof. The proof is straightforward by choosing $\epsilon = \frac{1}{c\gamma n_1 n_2}$ and substituting the expression for covering number in theorem 4.5.1 from lemma 4.11.7, and using the fact $2\log\left(\frac{2}{\alpha}\right) \leq 2(n_1 r + l)\log\left(\frac{2}{\alpha}\right)$ and $2l\log(\frac{n_2 re}{l}) \leq 2(n_1 r + l)\log(n_2 re)$ □

The above corollary reveals that the error measured in terms of per entry KL-divergence between the constrained maximum likelihood and the true matrix decays at the rate of γ^{-1} and is proportional to $n_1 r + l$. The quantity $n_1 r + l$ may be interpreted as the degree of freedom of the matrix set $\mathcal{X}$ in 4.8. For sparse factor models similar bounds for the error measured in terms per-entry square error for sparsity regularized maximum likelihood estimation from 1-bit quantized data was reported in [2] and [108].

In contrast to these, our analysis here is for multi-bit scenario and shows that similar rate is possible even with constrained maximum likelihood estimation. In section 4.7 we discuss a possible way to convert the bound in corollary 4.5.2 in terms of per-entry squared error.

4.6 Error guarantees based on the chaining principle

The error bounds obtained via covering number based approach was a relatively simple approach in which the supremum of the random variable $\frac{1}{2}\mathbb{E}U_{\mathbf{X}} - U_{\mathbf{X}}$ over the centers of an $\epsilon-$cover of the structured matrix set $\mathcal{X}$ is controlled, and the argument is extended to the entire set $\mathcal{X}$ by leveraging the Lipschitz continuity property. This covering number based approach involved taking union bound over the centers of the $\epsilon-$cover which resulted in log of covering number showing up in the upper bounds. This procedure is known to lead to loose bounds as the events over the centers of the ϵ-cover could be dependent for a dense enough cover. In such cases, an improvement in bounds is possible by following *multi-scale* approach in which the set is covered progressively by taking balls of increasing radius. Using this technique the supremum can be bounded by using the *Dudley's inequality* and sometimes it leads to an improvement in the upper bounds. In many cases, even *Dudley's inequality* fails to give tight enough bounds; then, one may follow the *chaining* principle [111,112]. In fact, for the Gaussian processes this approach gives optimal bounds. Motivated by this here we provide the error bounds obtained via the chaining principle.

In order to use the chaining principle we first define the following centered random variable indexed by $\mathbf{X} \in \mathcal{X}$

$$V_{\mathbf{X}} = \sum_{i,j} b_{ij} \log \frac{p(Z_{i,j}; X^*_{i,j})}{p(Z_{i,j}; X_{ij})} - \mathbb{E}\left(\sum_{i,j} b_{ij} \log \frac{p(Z_{i,j}; X^*_{i,j})}{p(Z_{i,j}; X_{ij})}\right).$$

The following lemma on sub-gaussian nature of increments $|V_{\mathbf{X}} - V_{\mathbf{Y}}|$ for $\mathbf{X}, \mathbf{Y} \in \mathcal{X}$ acts as the basic building block to obtain final error bounds using the chaining argument.

Lemma 4.6.1. *Suppose that $p(Z_{i,j}; X_{i,j}) \geq \delta_0$, for $\forall (i,j) \in [n_1] \times [n_2]$ and $\forall \mathbf{X} \in \mathcal{X}$*

(where $0 < \delta_0 < \frac{1}{2}$). The random variable for fixed $\mathbf{X} \in \mathcal{X}$

$$V_{\mathbf{X}} = \sum_{i,j} b_{ij} \log \frac{p(Z_{i,j}; X^*_{i,j})}{p(Z_{i,j}; X_{ij})} - \mathbb{E}\left(\sum_{i,j} b_{ij} \log \frac{p(Z_{i,j}; X^*_{i,j})}{p(Z_{i,j}; X_{ij})}\right)$$

satisfies the following concentration bound for any $\mathbf{Y} \in \mathcal{X}$

$$\textit{Prob}\left(|V_{\mathbf{X}} - V_{\mathbf{Y}}| \geq \tau d(\mathbf{X}, \mathbf{Y})\right) \leq 2e^{-\tau^2} \tag{4.9}$$

$d(\mathbf{X}, \mathbf{Y}) = L_g\sqrt{2n_1n_2}\|\mathbf{X} - \mathbf{Y}\|_F$ and L_g is defined as

$$L_g = \max_k \sup_{|t| \leq x_{max}} \left| \frac{f_W(\tau_k - t) - f_W(\tau_{k-1} - t)}{\int_{\tau_{k-1}-t}^{\tau_k - t} f_W(w)dw} \right|.$$

Proof. The proof is outlined in section 4.12. □

Using the above lemma the chaining principle can be used bound the supremum of random variables $V_{\mathbf{X}}$ over the set $\mathcal{X}$ in the following lemma.

Lemma 4.6.2. *Given a set of matrices $\mathcal{X} \subset \mathbb{R}^{n_1 \times n_2}$, assume $p(Z_{i,j}; X_{i,j}) \geq \delta_0$ for $\forall(i, j)$ and $\forall\mathbf{X} \in \mathcal{X}$, where $0 < \delta_0 < \frac{1}{2}$. The suprema of random variables*

$$V_{\mathbf{X}} = \sum_{i,j} b_{ij} \log \frac{p(Z_{i,j}; X^*_{i,j})}{p(Z_{i,j}; X_{ij})} - \mathbb{E}\left(\sum_{i,j} b_{ij} \log \frac{p(Z_{i,j}; X^*_{i,j})}{p(Z_{i,j}; X_{ij})}\right)$$

defined over $\mathbf{X} \in \mathcal{X}$ satisfies the following concentration bound

$$\textit{Prob}\left(\sup_{\mathbf{X} \in \mathcal{X}} |V_{\mathbf{X}}| \geq L_g\sqrt{n_1n_2}\left(C\gamma_2(\mathcal{X}, \ell_2) + \tau D\Delta_{\ell_2}(\mathcal{X})\right)\right) \leq e^{-\tau^2/2},$$

where C, D are absolute constants and

$$\gamma_2(\mathcal{X}, \ell_2) = \inf_{\mathcal{X}_a} \sup_{\mathbf{X} \in \mathcal{X}} \sum_{n=0}^{\infty} 2^{n/2} \inf_{\mathbf{Z} \in X_n} \|\mathbf{Z} - \mathbf{X}\|_F, \tag{4.10}$$

where $\mathcal{X}_a = (\mathcal{X}_n)_{n \geq 0}$ is a sequence of subset of $\mathcal{X}$, which satisfies $|\mathcal{X}_0| = 1$ and $|\mathcal{X}_n| \leq 2^{2^n}$ for all $n \geq 1$, and $\Delta_{\ell_2}(\mathcal{X}) = \sup_{\mathbf{X},\mathbf{Y} \in \mathcal{X}} \|\mathbf{X} - \mathbf{Y}\|_F$.

Proof. The proof is outline in section 4.13. □

Instantiating the above lemma for $\hat{\mathbf{X}}$ the performance bound based on the chaining argument can be obtained as stated in the following theorem.

Theorem 4.6.1. *Suppose $\mathcal{S}$ is chosen such that each entry is sampled independently with probability γ. Given the noise corrupted quantized observation of $\mathbf{X}^* \in \mathcal{X}$ taken as per* (4.1) *at subset of locations denoted by $\mathcal{S} \subseteq [n_1] \times [n_2]$. Assuming that the discrete pmf satisfies $p(Z_{i,j}; X_{i,j}) \geq \delta_0$ for $\forall (i,j)$ and $\forall \mathbf{X} \in \mathcal{X}$ $(0 < \delta_0 < \frac{1}{2})$, then with probability at least $1 - \alpha$ the constrained maximum likelihood estimate $\hat{\mathbf{X}}$ in* (4.3) *satisfies*

$$\frac{\mathrm{D}(p(\mathbf{Z};\mathbf{X}^*)||p(\mathbf{Z};\hat{\mathbf{X}}))}{n_1 n_2} \leq \frac{L_g\left[C\gamma_2(\mathcal{X},\ell_2) + \sqrt{2\log(\frac{1}{\alpha})}D\Delta_{\ell_2}(\mathcal{X})\right]}{\gamma\sqrt{n_1 n_2}} \tag{4.11}$$

where C, D are constants, $\Delta_{\ell_2}(\mathcal{X}) = \sup_{\mathbf{X},\mathbf{Y}\in\mathcal{X}} \|\mathbf{X}-\mathbf{Y}\|_F$, and L_g is defined as

$$L_g = max_k \sup_{|t|\leq x_{max}} \left|\frac{f_W(\tau_k - t) - f_W(\tau_{k-1} - t)}{\int_{\tau_{k-1}-t}^{\tau_k - t} f_W(w)dw}\right|.$$

Proof. Using the lemma 4.6.2 $\forall \mathbf{X} \in \mathcal{X}$ we have with probability at least $1 - e^{-\tau^2/2}$

$$\mathbb{E}\log\frac{p(\mathbf{Z}_\mathcal{S};\mathbf{X}^*_\mathcal{S})}{p(\mathbf{Z}_\mathcal{S};\mathbf{X}_\mathcal{S})} - \log\frac{p(\mathbf{Z}_\mathcal{S};\mathbf{X}^*_\mathcal{S})}{p(\mathbf{Z}_\mathcal{S};\mathbf{X}_\mathcal{S})} \leq L_g\sqrt{n_1 n_2}\left[C\gamma_2(\mathcal{X},\ell_2) + \sqrt{2\log\left(\frac{1}{\alpha}\right)}D\Delta_{\ell_2}(\mathcal{X})\right].$$

Since $\hat{\mathbf{X}} \in \mathcal{X}$ we instantiate the above inequality for $\hat{\mathbf{X}}$ and choose $\alpha = e^{-\tau^2/2}$, i.e., $\tau = \sqrt{2\log\left(\frac{1}{\alpha}\right)}$. So we have with probability at least $1 - \alpha$

$$\mathbb{E}\log\frac{p(\mathbf{Z}_\mathcal{S};\mathbf{X}^*_\mathcal{S})}{p(\mathbf{Z}_\mathcal{S};\hat{\mathbf{X}}_\mathcal{S})} \leq \log\frac{p(\mathbf{Z}_\mathcal{S};\mathbf{X}^*_\mathcal{S})}{p(\mathbf{Z}_\mathcal{S};\hat{\mathbf{X}}_\mathcal{S})} + L_g\sqrt{n_1 n_2}\left[C\gamma_2(\mathcal{X},\ell_2) + \sqrt{2\log\left(\frac{1}{\alpha}\right)}D\Delta_{\ell_2}(\mathcal{X})\right].$$

Now since $\hat{\mathbf{X}}$ maximizes $p(\mathbf{Z}_\mathcal{S};\mathbf{X}_\mathcal{S})$ over $\mathcal{X}$, and $\mathbf{X}^* \in \mathcal{X}$ we have $\log\frac{p(\mathbf{Z}_\mathcal{S};\mathbf{X}^*_\mathcal{S})}{p(\mathbf{Z}_\mathcal{S};\hat{\mathbf{X}}_\mathcal{S})} \leq 0$. Finally substituting $\mathbb{E}\log\frac{p(\mathbf{Z}_\mathcal{S};\tilde{\mathbf{X}}^*_\mathcal{S})}{p(\mathbf{Z}_\mathcal{S};\hat{\mathbf{X}}_\mathcal{S})} = \gamma\mathrm{D}(p(\mathbf{Z};\mathbf{X}^*)||p(\mathbf{Z};\hat{\mathbf{X}}))$ we have with probability atleast $1 - \alpha$,

$$\frac{\mathrm{D}(p(\mathbf{Z};\mathbf{X}^*)||p(\mathbf{Z};\hat{\mathbf{X}}))}{n_1 n_2} \leq \frac{L_g\left[C\gamma_2(\mathcal{X},\ell_2) + \sqrt{2\log\left(\frac{1}{\alpha}\right)}D\Delta_{\ell_2}(\mathcal{X})\right]}{\gamma\sqrt{n_1 n_2}}.$$

□

The theorem 4.6.1 bounds the per-entry KL-divergence in terms of $\gamma_2(\mathcal{X},\ell_2)$ and $\Delta_{\ell_2}(\mathcal{X})$. These two constants are purely the geometric property of the set $\mathcal{X}$. The term $\Delta_{\ell_2}(\mathcal{X})$ can be interpreted upper bounded as $\Delta_{\ell_2}(\mathcal{X}) \leq 2\max_{\mathbf{X}\in\mathcal{X}} \|\mathbf{X}\|_F$. The $\gamma_2(\mathcal{X},\ell_2)$ and $\Delta_{\ell_2}(\mathcal{X})$ could interpreted as the complexity of set $\mathcal{X}$ and similar to the bound in theorem 4.5.1 the bounds in theorem 4.6.1 reveals the γ^{-1} rate. However, in contrast to the bound in theorem 4.6.1 obtained does not have a constant $c = \frac{1}{4\frac{1-\delta_0}{\delta_0}(\log\frac{1-\delta_0}{\delta_0}+1)}$ which could be very large for small values δ_0. The term $\gamma_2(\mathcal{X}, \ell_2)$ is non-trivial to calculate however it can be upper bounded using known inequalities or can be shown to proportional to a quantity which is easy to calculate. In particular, the following relationship is quite useful for obtaining interpretable bounds [111].

$$\gamma_2(\mathcal{X}, \ell_2) \leq \beta \cdot \Omega(\mathcal{X}), \tag{4.12}$$

where β is a constant and $\Omega(\mathcal{X})$ is the Gaussian width of the set $\mathcal{X}$ defined as follows

$$\Omega(\mathcal{X}) = \mathbb{E}\left[\sup_{\mathbf{X}\in\mathcal{X}} \mathrm{Tr}\left(\mathbf{G}\mathbf{X}\right)\right], \tag{4.13}$$

where $\mathbf{G}$ is a Gaussian random matrix of size $n_2 \times n_1$ with iid entries following standard Gaussian distribution $\mathcal{N}(0,1)$. The Gaussian width is of the set $\mathcal{X}$ is a measure of complexity of the set $\mathcal{X}$ akin to the degrees of freedom [66]. Next we instantiate the theorem 4.6.1 for a specific set of matrices to demonstrate its utility.

4.6.1 Set of low rank matrices

Here we instantiate theorem 4.6.1 for the set of low rank matrices defined as follows

$$\mathcal{X} = \left\{\sum_{i=1}^{r} \mathbf{d}_i\mathbf{a}_i^T \middle| \mathbf{d}_i \in \mathbb{R}^{n_1}, \|\mathbf{d}_i\|_2 \leq 1, \mathbf{a}_i \in \mathbb{R}^{n_2}, \|\mathbf{a}_i\|_2 \leq 1, i \in [r]\right\}, \tag{4.14}$$

where $r < \min\{n_1, n_2\}$. Note that the matrices in set $\mathcal{X}$ are not explicitly bounded in magnitude however using the definition of the set it is possible to obtain x_{max}.

Corollary 4.6.1. *Under the assumptions of theorem 4.6.1. For the set of low rank matrices $\mathcal{X}$ defined in* (4.14) *given the noise corrupted quantized observation of $\mathbf{X}^* \in \mathcal{X}$ taken as per* (4.1) *then with probability at least $1-\alpha$ the constrained maximum likelihood estimate $\hat{\mathbf{X}}$ in* (4.3) *satisfies*

$$\frac{\mathrm{D}(p(\mathbf{Z};\mathbf{X}^*)||p(\mathbf{Z};\hat{\mathbf{X}}))}{n_1n_2} \leq \frac{L_g\left[C\beta + \sqrt{2\log\left(\frac{1}{\alpha}\right)}D\right] r(n_1+n_2)}{\gamma\sqrt{n_1n_2}} \tag{4.15}$$

where C, D, β are constants and L_g as defined in theorem 4.6.1.

Proof. The proof is outlined in section 4.14. □

The bound in the above corollary as compared to the bound in corollary 4.5.1 decays with similar as far as decay with the sampling rate γ is concerned however above bound does not suffer from the constant c which could be very large for small values of δ_0. The decay rate in corollary 4.6.1 is still better than the ones reported in [98, 104] and our analysis here holds for multiple bit scenarios as well. Next we discuss a possible way to convert the bound in corollary 4.6.1 in terms of per-entry squared error.

4.7 Further discussion on the error bounds

The bounds the corollaries 4.5.1,4.5.2, 4.6.1 suggest that the average per entry KL divergence between the pmfs parametrizes by the estimate $\hat{\mathbf{X}}$ and $\hat{\mathbf{X}}^*$ error decays reciprocally with respect to sampling probability γ. The KL divergence is the measure of distance between the pmfs however nearness in pmfs also imply nearness in the parameter space. However, this depends on the specific noise density. In particular, if there is a constant L_f such that we have

$$L_f \|\mathbf{X}^* - \hat{\mathbf{X}}\|_F^2 \leq D(p(\mathbf{Z}; \mathbf{X}^*) || p(\mathbf{Z}; \hat{\mathbf{X}})) \tag{4.16}$$

then the error bounds in corollaries 4.5.1, 4.5.2, 4.6.1 can be converted to per-entry estimation error guarantees measured in squared error similar to the ones obtained in [98,104,107]. Further, the dependence on K is not obvious in the above error bounds. The error bounds depend on K via the constant L_g and L_f. These constants are a function of shape of the noise pdf $f_W(W)$ and thresholds $\{\tau_i\}_{i=0}^K$.

4.8 Numerical algorithms

The constrained maximum log likelihood estimates in (4.3) can be written as

$$\widehat{\mathbf{X}} = \arg\min_{\mathbf{X} \in \mathcal{X}} \sum_{i,j} b_{i,j} \ell(Z_{i,j}, X_{i,j}), ,$$

where we have used the shorthand notation $\ell(z,x) = -\log(p(z;x)$ to denote the negative log likelihood function of noise density $f_W(w)$ given by

$$\ell(z,x) = -\sum_{k=1}^{K} \mathbf{1}\left(z=k\right)\log\left(\int_{\tau_{k-1}-x}^{\tau_k-x} f_W(w)dw,\right). \tag{4.17}$$

It is known that $-\log\left(\int_{\tau_{k-1}-v}^{\tau_k-v} f_W(w)dw\right)$ is a convex function of v for all log-concave densities [113] under such densities the objective function for the above optimization problem is a convex function. However, depending on the specific matrix sets $\mathcal{X}$ the problem may be convex or non-convex.

4.8.1 For convex $\mathcal{X}$

For convex $\mathcal{X}$ the overall problem is convex and can be solved via projected gradient descent for the differentiable $\ell(\cdot)$ otherwise the projected sub-gradient method may be used. The overall algorithm is summarized as Algorithm 5. The algorithm is described in terms of sub-gradient whenever the $\ell(\cdot)$ is differentiable the projected sub-gradient method reduces to projected gradient. Notice that the algorithm uses diminishing step sizes which guarantees convergence to the solution of the problem.

Algorithm 5 Projected sub-gradient Algorithm to solve $\min_{\mathbf{X}\in\mathcal{X}} \sum_{i,j} b_{i,j}\ell(Z_{i,j};X_{i,j})$

```
Initialize: $\mathbf{X}^{(0)} = \mathbf{0}$
  Repeat for $t = 1, 2, \cdots$ until convergence.
    $\mathbf{U}^{(t-1)} \in \partial\left(\sum_{i,j} b_{i,j}\ell(Z_{i,j};X_{i,j}^{(t-1)})\right)$
    $\mathbf{X}^{(t)} = \arg\min_{\mathbf{X}\in\mathcal{X}} \|\mathbf{X} - \left(\mathbf{X}^{(t-1)} - \frac{\alpha}{t}\mathbf{U}^{(t-1)}\right)\|_F^2$
Output: $\mathbf{X} = \mathbf{X}^{(t)}$
```

4.8.2 For non-convex $\mathcal{X}$

For non-convex sets the optimization algorithm vary a lot depening on the specific set $\mathcal{X}$. Here we focus on factor models in which each matrix $\mathbf{X} \in \mathcal{X}$ can be decomposed as $\mathbf{X} = \mathbf{DA}$ where $\mathbf{D} \in \mathcal{D}$ and $\mathbf{A} \in \mathcal{A}$. For such sets the constrained maximum likelihood problem takes the following form.

$$\begin{aligned} \min_{\mathbf{D}\in\mathbb{R}^{n_1\times r},\mathbf{A}\in\mathbb{R}^{r\times n_2}} \quad & \sum_{i,j} b_{i,j}\ell(Z_{i,j},X_{i,j}) + I_{\mathcal{D}}(\mathbf{D}) + I_{\mathcal{A}}(\mathbf{A}) \\ \text{s.t.} \quad & \mathbf{X} = \mathbf{DA}. \end{aligned} \tag{4.18}$$

where $I_{\mathcal{D}}(\cdot)$, and $I_{\mathcal{A}}(\cdot)$ are the indicator functions of the sets $\mathcal{D}$ and $\mathcal{A}$ respectively.[1]

Motivated by the Alternating Direction Method of Multipliers (ADMM) approach proposed for matrix completion for sparse factor models in [2] here we extend it to quantized setting for general factor models. The augmented Lagrangian of (4.18) as

$$\mathcal{L}(\mathbf{D},\mathbf{A},\mathbf{X},\mathbf{\Lambda}) = \sum_{i,j} b_{i,j}\ell(Z_{i,j},X_{i,j}) + I_{\mathcal{D}}(\mathbf{D}) + I_{\mathcal{A}}(\mathbf{A}) + \operatorname{tr}\left(\mathbf{\Lambda}(\mathbf{X}-\mathbf{DA})\right) + \frac{\rho}{2}\|\mathbf{X}-\mathbf{DA}\|_{\mathrm{F}}^2,$$

where $\mathbf{\Lambda}$ is a matrix of Lagrange multipliers and $\rho > 0$.Starting with some feasible $\mathbf{A}^{(0)}, \mathbf{D}^{(0)}, \mathbf{\Lambda}^{(0)}$ we iteratively update $\mathbf{X}$, $\mathbf{A}$, $\mathbf{D}$, and $\mathbf{\Lambda}$ according to

$$\begin{aligned} (\mathbf{S1}:)\ \mathbf{X}^{(k+1)} &:= \arg\min_{\mathbf{X}\in\mathbb{R}^{n_1\times n_2}} \mathcal{L}(\mathbf{D}^{(k)},\mathbf{A}^{(k)},\mathbf{X},\mathbf{\Lambda}^{(k)}) \\ (\mathbf{S2}:)\ \mathbf{A}^{(k+1)} &:= \arg\min_{\mathbf{A}\in\mathbb{R}^{r\times n_2}} \mathcal{L}(\mathbf{D}^{(k)},\mathbf{A},\mathbf{X}^{(k+1)},\mathbf{\Lambda}^{(k)}) \\ (\mathbf{S3}:)\ \mathbf{D}^{(k+1)} &:= \arg\min_{\mathbf{D}\in\mathbb{R}^{n_1\times r}} \mathcal{L}(\mathbf{D},\mathbf{A}^{(k+1)},\mathbf{X}^{(k+1)},\mathbf{\Lambda}^{(k)}) \\ (\mathbf{S4}:)\ \mathbf{\Lambda}^{(k+1)} &= \mathbf{\Lambda}^{(k)} + \rho(\mathbf{X}^{(k+1)} - \mathbf{D}^{(k+1)}\mathbf{A}^{(k+1)}), \end{aligned}$$

until convergence, which is based on that the norms of primal and dual residuals become sufficiently small (as described in [114]). After completing the square and ignoring constant terms problem the step **S1** is equivalent to the following problem

$$\min_{\mathbf{X}\in\mathbb{R}^{n_1\times n_2}} \quad \sum_{i,j} b_{i,j}\ell(Z_{i,j},X_{i,j}) + \frac{\rho}{2}\left\|\mathbf{X} - \mathbf{D}^{(k)}\mathbf{A}^{(k)} + \frac{\mathbf{\Lambda}^{(k)}}{\rho}\right\|_F^2.$$

The above problem is separable in each entry $X_{i,j}$ and the entries can be updated in parallel by solving the scalar convex optimization problem for each entry as follows

$$X_{i,j}^{(k+1)} = \begin{cases} \arg\min_x \ell(Z_{i,j},x) + \frac{\rho}{2}\left(x - (\mathbf{D}^{(k)}\mathbf{A}^{(k)})_{i,j} + \frac{(\mathbf{\Lambda}^{(k)})_{i,j}}{\rho}\right)^2, & \text{if } b_{i,j} = 1 \\ (\mathbf{D}^{(k)}\mathbf{A}^{(k)})_{i,j} - \frac{(\mathbf{\Lambda}^{(k)})_{i,j}}{\rho}, & \text{otherwise} \end{cases}$$

[1] The indicator function takes value 0 if its argument is an element of the set described as the subscript otherwise it takes value ∞.

Algorithm 6 ADMM algorithm for solving problem (4.18)

```
Inputs: $\epsilon_1, \epsilon_2,\ \Delta_1,\ \Delta_2,\ \Delta_1^{\text{stop}},\ \Delta_2^{\text{stop}},\ \eta,\ \rho^{(0)} > 0$
Initialize: $\mathbf{D}^{(0)} \in \mathcal{D}$ , $\mathbf{A}^{(0)} \in \mathcal{A}$, $\mathbf{\Lambda}^{(0)}$.
  repeat
    $X_{i,j}^{(k+1)} = \begin{cases} \arg\min_x \ell(Z_{i,j}, x) + \frac{\rho}{2}\left(x - (\mathbf{D}^{(k)}\mathbf{A}^{(k)})_{i,j} + \frac{(\mathbf{\Lambda}^{(k)})_{i,j}}{\rho}\right)^2, & \text{if } b_{i,j} = 1 \\ (\mathbf{D}^{(k)}\mathbf{A}^{(k)})_{i,j} - \frac{(\mathbf{\Lambda}^{(k)})_{i,j}}{\rho}, & \text{ig } b_{i,j} = 0 \end{cases}$
    $\mathbf{A}^{(k+1)} := \arg\min_{\mathbf{A} \in \mathbb{R}^{r \times n_2}} I_{\mathcal{A}}(\mathbf{A}) + \frac{\rho}{2}\left\|\mathbf{X}^{(k+1)} - \mathbf{D}^{(k)}\mathbf{A} + \frac{\mathbf{\Lambda}^{(k)}}{\rho}\right\|_F^2$
    $\mathbf{D}^{(k+1)} = \arg\min_{\mathbf{D} \in \mathbb{R}^{n_1 \times r}} I_{\mathcal{D}}(\mathbf{D}) + \frac{\rho}{2}\left\|\mathbf{X}^{(k+1)} - \mathbf{D}\mathbf{A}^{(k+1)} + \frac{\mathbf{\Lambda}^{(k)}}{\rho}\right\|_F^2$
    $\mathbf{\Lambda}^{(k+1)} = \mathbf{\Lambda}^{(k)} + \rho^{(k)}(\mathbf{X}^{(k+1)} - \mathbf{D}^{(k+1)}\mathbf{A}^{(k+1)})$
    Set $\Delta_1 = \|\mathbf{X}^{(k+1)} - \mathbf{D}^{(k+1)}\mathbf{A}^{(k+1)}\|_F$ and $\Delta_2 = \rho^{(k)} \cdot \|\mathbf{D}^{(k)}\mathbf{A}^{(k)} - \mathbf{D}^{(k+1)}\mathbf{A}^{(k+1)}\|_F$
    $\rho^{(k+1)} = \begin{cases} \eta \cdot \rho^{(k)}, & \text{if } \Delta_1 \geq 10 \cdot \Delta_2 \\ \rho^{(k)}/\eta, & \text{if } \Delta_2 \geq 10 \cdot \Delta_1 \\ \rho^{(k)}, & \text{otherwise} \end{cases}$
  until $\Delta_1 \leq \Delta_1^{\text{stop}}$ and $\Delta_2 \leq \Delta_2^{\text{stop}}$
Output: $\mathbf{D} = \mathbf{D}^{(k+1)}$ and $\mathbf{A} = \mathbf{A}^{(k+1)}$
```

For convex $\ell(\cdot)$ the subproblem for $b_{i,j} = 1$ is a strictly convex problem which can be solved via first order methods like gradient descent for differentiable $\ell(\cdot)$, via sub-gradient method for non-differentiable $\ell(\cdot)$, and for double differentiable $\ell(\cdot)$ one may even employ second order Newton type methods. Further, again completing the square and ignoring the constant terms the subproblem **S2** is equivalent to

$$\mathbf{A}^{(k+1)} = \arg \min_{\mathbf{A} \in \mathbb{R}^{r \times n_2}} \quad I_{\mathcal{A}}(\mathbf{A}) + \frac{\rho}{2}\left\|\mathbf{X}^{(k+1)} - \mathbf{D}^{(k)}\mathbf{A} + \frac{\mathbf{\Lambda}^{(k)}}{\rho}\right\|_F^2 .$$

The above problem is minimization of a convex smooth function over the constrained set $\mathcal{A}$. For convex $\mathcal{A}$, one may employ projected sub-gradient methods similar to Algorithm 5. For non-convex $\mathcal{A}$, the above step changes significantly from one set to another. A general algorithm is not possible. For the set of sparse matrices, which are one of the main focus of this paper, we employ iterative hard thresholding (IHT) type algorithm to solve it [115]. Finally, after completing the square and ignoring the constant terms, we see that the subproblem **S3** is equivalent to

$$\mathbf{D}^{(k+1)} = \arg \min_{\mathbf{D} \in \mathbb{R}^{n_1 \times r}} \quad I_{\mathcal{D}}(\mathbf{D}) + \frac{\rho}{2} \left\| \mathbf{X}^{(k+1)} - \mathbf{D}\mathbf{A}^{(k+1)} + \frac{\mathbf{\Lambda}^{(k)}}{\rho} \right\|_F^2 . \tag{4.19}$$

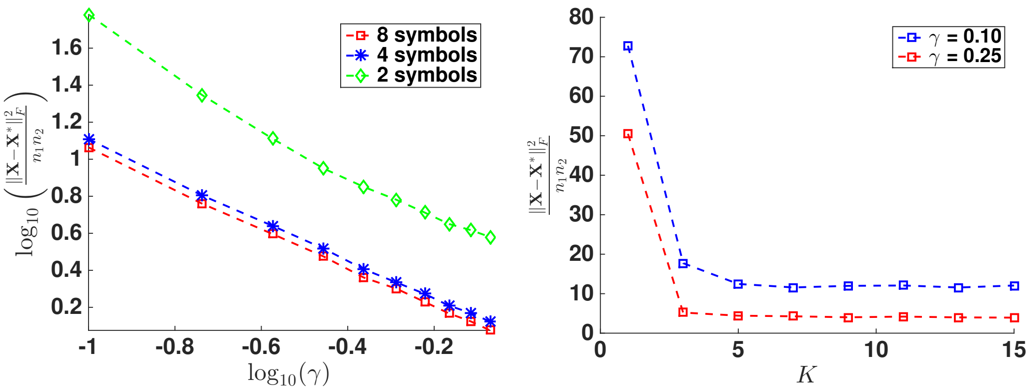
For convex $\mathcal{D}$ one may a employ a projected gradient method similar to Algorithm 5 whereas for non-convex $\mathcal{D}$ the algorithm depends on the specific set. Our overall algorithmic approach is summarized in Algorithm 6. Note that in final algorithm the value of ρ is varied for faster convergence as suggested in [114].

4.9 Experimental evaluation

We perform experimental validation on synthetic data set for sparse factor models. The true data matrix for the experiment was generated randomly by first generating the matrices $\mathbf{D}^* \in \mathbb{R}^{1000\times 5}$, $\mathbf{A}^* \in \mathbb{R}^{5\times 1000}$ and then using these to obtain the true matrix as $\mathbf{X}^* = \mathbf{D}^*\mathbf{A}^*$. We randomly generate $\mathbf{D}^*$ such its entries i.i.d. Gaussian distribution with zero mean and variance 4 and the threshold its entries to lie in the range $[-1, 1]$. Similarly, we generate random column sparse $\mathbf{A}^*$ by first randomly selecting $s = 2$ non-zero locations in each column, and equating them i.i.d. Gaussian random numbers with zero mean and variance 400 and thresholding these to the range $[-10, 10]$. Using these random $\mathbf{D}^*, \mathbf{A}^*$ we obtain the matrix $\mathbf{D}^*\mathbf{A}^*$ and threshold its entries to the range $[-20, 20]$ to obtain $\mathbf{X}^*$. The matrix $\mathbf{X}^*$ generated as above is observed at subset of its entries $\mathcal{S}$ by observing each entry with probability γ in a i.i.d. fashion to obtained quantized observations in noise as follows

$$Y_{ij} = \mathcal{Q}(X^*_{ij} + W_{ij}), \quad (i, j) \in \mathcal{S},$$

where $W_{i,j}$ are i.i.d. Gaussian noise with zero mean and variance $\sigma^2 = 100$, and the quantizer $\mathcal{Q}(\cdot)$ maps observations to K symbols by choosing the thresholds $\{\tau_i\}_{i=1}^{K-1}$ by uniformly dividing the interval $[-20, 20]$ in $K - 2$ parts and $\tau_0 = -\infty$ and $\tau_K = +\infty$.



(a) (b)

Figure 4.2: Results of synthetic experiments for quantized matrix completion under sparse factor models. The panel (a) contains log per-entry squared error $\log\left(\frac{\|\hat{\mathbf{X}}-\mathbf{X}^*\|_F^2}{n_1 n_2}\right)$ vs. log sampling rate $\log_{10}(\gamma)$ plots for $K = 8, 4, 2$. Panel (b) contains per-entry squared error $\frac{\|\mathbf{X}-\mathbf{X}^*\|_F^2}{n_1 n_2}$ vs. number of symbols in quantizer K plots for $\gamma = 0.10, 0.25$.

Using these quantized observations, the constrained maximum likelihood optimization problem is solved to obtain the estimate $\hat{\mathbf{X}}$. For this we use Algorithm 6 in which for $\mathbf{A}$ update step iterative hard thresholding algorithm is employed whereas $\mathbf{D}$ is updated via projected gradient descent algorithm. We repeated this experiment for various values of sampling rates γ and number of symbols K. The final results are shown in figure 4.2 wherein panel (a) we plot the $\log\left(\frac{\|\hat{\mathbf{X}}-\mathbf{X}^*\|_F^2}{(n_1 n_2)}\right)$ vs. $\log(\gamma)$ for $K = 8, 4, 2$. The plots for each value of K have the slope of about -1. This suggests that the error measured even in terms of per-entry squared loss decays reciprocally with the sampling rate γ and suggests the existence of a constant L_f as described in (4.16). Further, also observe that curve shifts down as we increase the number of symbols from 2 to 8. This implies that the error decreases as we increase the number of symbols K. However, an interesting thing to note here is that as we increase the number of symbols from 2 to 4 there is large shift in the curve as compared to when we increase the number of symbols from 4 to 8. This observation is further confirmed by plot in panel (b) of figure 4.2. This empirical observation suggest a *law of diminishing return* type dependence on the number of symbols K.

4.10 Summary

We studied the problem of matrix completion arising in noisy and quantized setting. Following up on initial work on noisy matrix completion for sparse factor models we investigated the general problem of quantized matrix completion for structured matrix sets and obtained two generic theorems which were specialized for specific matrix sets. We also provided generic algorithmic framework for solving the quantized matrix completion which was used to empirically verify the bounds obtained for the sparse factor models. See section 7.3 for more discussion on possible future directions of research.

4.11 Appendix

4.11.1 Useful lemmata

Lemma 4.11.1. *Suppose two pmfs p, q of discrete random variable Z satisfy $0 < \delta_0 \leq p(Z = k), q(Z = k)$ for all $k = 1, \cdots, K$ then we have*

$$\frac{\sum_{k=1}^{K}(p(Z=k)-q(Z=k))^2}{2(1-\delta_0)} \leq \mathrm{D}(p\|q) \leq \frac{\sum_{k=1}^{K}(p(Z=k)-q(Z=k))^2}{2\delta_0} \tag{4.20}$$

Proof. Let $q(Z = k) = \pi_k$ and define $\delta_k := p(Z = k) - q(Z = k)$ we can express $p(Z = k)$ as $p(Z = k) = \pi_k + \delta_k$. Since $\sum_{k=1}^{K} p(Z = k) = \sum_{k=1}^{K} q(Z = k) = 1$ the δ_k's satisfy

$$\delta_0 \leq \pi_k + \delta_k, \forall k = 1, \cdots, K \text{ and } \sum_{k=1}^{K} \delta_k = 0. \tag{4.21}$$

The KL-Divergence $D(p\|q)$ can be written as the function $g(\underline{\delta})$ of vector of density differences $\underline{\delta} = [\delta_1, \cdots, \delta_K]^T$.

$$g(\underline{\delta}) = \sum_{k=1}^{K} (\pi_k + \delta_k) \log\left(\frac{\pi_k + \delta_k}{\pi_k}\right)$$

The gradient and hessian of $g(\underline{\delta})$ w.r.t $\underline{\delta}$ are easily shown to be

$$\begin{aligned} \nabla g(\underline{\delta}) &= \left[1 + \log\left(\tfrac{\pi_1+\delta_1}{\pi_1}\right), \cdots, 1 + \log\left(\tfrac{\pi_K+\delta_K}{\pi_K}\right)\right]^T, \\ \nabla^2 g(\underline{\delta}) &= \mathrm{Diag}\left(\left[\tfrac{1}{\pi_1+\delta_1}, \cdots, \tfrac{1}{\pi_K+\delta_K}\right]^T\right), \end{aligned}$$

where Diag$(\cdot)$ is a diagonal matrix from a vector in such a manner that the i^{th} component is the i^{th} diagonal element of the resultant matrix. Since $0 < \delta_0 \leq p(Z = k), q(Z = k)$ it implies that $p(Z = k), q(Z = k) \leq 1 - \delta_0 < 1$ we have

$$\mathbf{I}/(1-\delta_0) \preceq \nabla^2 g(\underline{\delta}) \preceq \mathbf{I}/\delta_0. \tag{4.22}$$

Next we use the Taylor's expansion of variance of $g(\underline{\delta})$ around zero vector to get

$$g(\underline{\delta}) = g(\mathbf{0}) + \nabla g(\mathbf{0})^T \underline{\delta} + \frac{1}{2}\underline{\delta}^T \nabla^2 g(t\underline{\delta})\underline{\delta}, \text{ for some } t \in [0,1]$$

Note that if $\underline{\delta}$ is a valid then $t\underline{\delta}$ for $t \in [0,1]$ is also valid as it satisfies both the conditions $\delta_0 \leq \pi_k + t\delta_k, \forall k = 1, \cdots, K$ and $\sum_{k=1}^K t\delta_k = 0$. Now, using the bound on hessian in (4.22) and substituting it in Taylor series expansion we have

$$g(\mathbf{0}) + \nabla g(\mathbf{0})^T \underline{\delta} + \frac{\underline{\delta}^T \mathbf{I}_n \underline{\delta}}{2(1-\delta_0)} \leq g(\underline{\delta}) \leq g(\mathbf{0}) + \nabla g(\mathbf{0})^T \underline{\delta} + \frac{\underline{\delta}^T \mathbf{I}_n \underline{\delta}}{2\delta_0}$$

it is straightforward to see that $g(\mathbf{0}) = 0$ and $\nabla g(\mathbf{0}) = [1, \cdots, 1]^T$ so we have

$$\sum_{k=1}^K \delta_k + \frac{\sum_{k=1}^K \delta_k^2}{2(1-\delta_0)} \leq g(\underline{\delta}) \leq \sum_{k=1}^K \delta_k + \frac{\sum_{k=1}^K \delta_k^2}{2\delta_0},$$

But, by (4.21) we have

$$\frac{\sum_{k=1}^K (p(Z=k) - q(Z=k))^2}{2(1-\delta_0)} \leq \mathrm{D}(p\|q) \leq \frac{\sum_{k=1}^K (p(Z=k) - q(Z=k))^2}{2\delta_0}$$

□

Lemma 4.11.2. *Suppose two pmfs p, q of discrete random variable Z satisfy $0 < \delta_0 \leq p(Z = k), q(Z = k) < 1$ for all $k = 1, \cdots, K$ and an independent Bernoulli random variable $b \sim Ber(\gamma)$ then we have*

$$\mathrm{var}_{Z \sim p, b \sim Ber(\gamma)}\left(b \log \frac{p(Z)}{q(Z)}\right) \leq \frac{2(1-\delta_0)}{\delta_0}\gamma \mathrm{D}(p\|q)$$

Proof. Notational brevity we denote the random variable $b \log \frac{p(Z)}{q(Z)}$ by U. So the variance of U is given by

$$
\begin{aligned}
\mathrm{var}_{Z\sim p,b}(U) &= \mathbb{E}_{Z\sim p,b}\left(U^2\right) - \left[\mathbb{E}_{Z\sim p,b}\left(U\right)\right]^2 \\
&= \mathbb{E}_{Z\sim p,b}\left(b^2 \log^2 \frac{p(Z)}{q(Z)}\right) - \left[\mathbb{E}_{Z\sim p,b}\left(b \log \frac{p(Z)}{q(Z)}\right)\right]^2 \\
&= \mathbb{E}_b(b^2)\mathbb{E}_{Z\sim p}\left(\log^2 \frac{p(Z)}{q(Z)}\right) - \left[\mathbb{E}_b(b)\mathbb{E}_{Z\sim p}\left(\log \frac{p(Z)}{q(Z)}\right)\right]^2 \\
&= \gamma\mathbb{E}_{Z\sim p}\left(\log^2 \frac{p(Z)}{q(Z)}\right) - \gamma^2\left[\mathbb{E}_{Z\sim p}\left(\log \frac{p(Z)}{q(Z)}\right)\right]^2 \qquad (4.23)
\end{aligned}
$$

Let $q(Z = k) = \pi_k$ and define $\delta_k := p(Z = k) - q(Z = k)$ we can express $p(Z = k)$ as $p(Z = k) = \pi_k + \delta_k$. Now the variance of U can be written as the function $g(\underline{\delta})$ of vector of density differences $\underline{\delta} = [\delta_1, \cdots, \delta_K]^T$

$$
\begin{aligned}
\mathrm{var}_{Z\sim p,b}(U) &= g(\underline{\delta}) \\
&= \gamma\sum_{i=1}^{K}(\pi_i + \delta_i)\log^2\frac{\pi_i + \delta_i}{\pi_i} - \gamma^2\left(\sum_{i=1}^{K}(\pi_i + \delta_i)\log\frac{\pi_i + \delta_i}{\pi_i}\right)^2
\end{aligned}
$$

Since $\underline{\delta}$ is vector of differences of valid pmfs p, q it must satisfy the following properties

$$
\delta_0 < \pi_i + \delta_i, \forall i = 1, \cdots, K \text{ and } \sum_{i=1}^{K}\delta_i = 0
$$

Next we find gradient of variance $g(\underline{\delta})$ w.r.t $\underline{\delta}$. The partial derivative w.r.t δ_k is given by

$$
\begin{aligned}
\frac{\partial g(\underline{\delta})}{\partial \delta_k} =& \gamma\left[\log^2\frac{\pi_k + \delta_k}{\pi_k} + 2\log\left(\frac{\pi_k + \delta_k}{\pi_k}\right)\right] \\
&- 2\gamma^2\left[\sum_{i=1}^{K}(\pi_i + \delta_i)\log\frac{\pi_i + \delta_i}{\pi_i}\right]\left[\log\frac{\pi_k + \delta_k}{\pi_k} + 1\right]
\end{aligned}
$$

For Hessian we need second order derivatives it is easy to see that for $k \neq l$

$$
\frac{\partial^2 g(\underline{\delta})}{\partial \delta_k \partial \delta_l} = -2\gamma^2\left[\log\frac{\pi_k + \delta_k}{\pi_k} + 1\right]\left[\log\frac{\pi_l + \delta_l}{\pi_l} + 1\right], \quad \text{for } k \neq l
$$

and for $k = l$ we have

$$
\begin{aligned}
\frac{\partial^2 g(\underline{\delta})}{\partial^2 \delta_k} = \frac{2\gamma}{\pi_k + \delta_k}\left[\log\frac{\pi_k + \delta_k}{\pi_k} + 1\right] - 2\gamma^2\left[\log\frac{\pi_k + \delta_k}{\pi_k} + 1\right]^2 \\
- \frac{2\gamma^2}{\pi_k + \delta_k}\left[\sum_{i=1}^{K}(\pi_i + \delta_i)\log\frac{\pi_i + \delta_i}{\pi_i}\right]
\end{aligned}
$$

Finally Hessian can be written as

$$\begin{aligned}\nabla^2 g(\underline{\delta}) =&\text{Diag}\left(\left[\frac{2\gamma}{\pi_1+\delta_1}\left(\log\frac{\pi_1+\delta_1}{\pi_1}+1\right),\cdots,\frac{2\gamma}{\pi_K+\delta_K}\left(\log\frac{\pi_K+\delta_K}{\pi_K}+1\right)\right]\right)\\ &-2\gamma^2\begin{bmatrix}\log\frac{\pi_1+\delta_1}{\pi_1}+1\\ \vdots\\ \log\frac{\pi_K+\delta_K}{\pi_K}+1\end{bmatrix}\begin{bmatrix}\log\frac{\pi_1+\delta_1}{\pi_1}+1, & \cdots, & \log\frac{\pi_K+\delta_K}{\pi_K}+1\end{bmatrix}\\ &-\left[\sum_{k=1}^{K}(\pi_k+\delta_k)\log\frac{\pi_k+\delta_k}{\pi_k}\right]\text{Diag}\left(\left[\frac{2\gamma^2}{\pi_1+\delta_1},\cdots,\frac{2\gamma^2}{\pi_K+\delta_K}\right]\right)\\ &\preceq\text{Diag}\left(\left[\frac{2\gamma}{\pi_1+\delta_1}\left(\log\frac{\pi_1+\delta_1}{\pi_1}+1\right),\cdots,\frac{2\gamma}{\pi_K+\delta_K}\left(\log\frac{\pi_K+\delta_K}{\pi_K}+1\right)\right]\right)\end{aligned}$$

The model assumptions $0<\delta_0\le p(Z=k),q(Z=k)$ implies that $p(Z=k),q(Z=k)\le 1-\delta_0$ for all $k=1,\cdots,K$. Using this we have

$$\begin{aligned}&\text{Diag}\left(\left[\frac{2\gamma}{\pi_1+\delta_1}\left(\log\frac{\pi_1+\delta_1}{\pi_1}+1\right),\cdots,\frac{2\gamma}{\pi_K+\delta_K}\left(\log\frac{\pi_K+\delta_K}{\pi_K}+1\right)\right]\right)\\ &\quad\preceq\text{Diag}\left(\left[\frac{2\gamma}{\pi_1},\cdots,\frac{2\gamma}{\pi_K}\right]\right)\preceq\frac{2\gamma}{\delta_0}\mathbf{I},\end{aligned}$$

where we have used the fact that $\log x\le x-1$ and $\pi_k\ge\delta_0$. Next we have the following simple relation

$$\nabla^2 g(\underline{\delta})\preceq\frac{2\gamma}{\delta_0}\mathbf{I}$$

for all valid $\underline{\delta}=[\delta_1,\cdots,\delta_K]^T$ obtained by difference of pmf values. Now using Taylor's expansion of variance of U around zero vector is given by

$$g(\underline{\delta})=g(\mathbf{0})+\nabla g(\mathbf{0})^T\underline{\delta}+\frac{1}{2}\underline{\delta}^T\nabla^2 g(t\underline{\delta})\underline{\delta},\ \text{for some } t\in[0,1]$$

Note that if $\underline{\delta}$ is a valid then $t\underline{\delta}$ for $t\in[0,1]$ is also valid as it satisfies both the conditions in (4.24). Further, it is straightforward to see that $g(\mathbf{0})=0$ and $\nabla g(\mathbf{0})=\mathbf{0}$ so we have

$$\begin{aligned}g(\underline{\delta})&=\frac{1}{2}\underline{\delta}^T\nabla^2 g(t\underline{\delta})\underline{\delta},\ \text{for some } t\in[0,1]\\ &\le\frac{\gamma}{\delta_0}\|\underline{\delta}\|^2=\frac{\gamma}{\delta_0}\left[\sum_{i=1}^{K}(p(Z=i)-q(Z=i))^2\right]\end{aligned}$$

Further using Lemma 4.11.1 we can upper bound $\sum_{i=1}^{K} \left(p(Z=i) - q(Z=i)\right)^2$ as

$$\sum_{i=1}^{K} \left(p(Z=i) - q(Z=i)\right)^2 \leq 2(1-\delta_0)\mathrm{D}(p\|q)$$

So finally we have

$$g(\underline{\delta}) \leq \frac{2(1-\delta_0)}{\delta_0}\gamma \mathrm{D}(p\|q)$$

□

Lemma 4.11.3. *Assuming that the function* $g_k(x) = \log\left(\int_{\tau_{k-1}-x}^{\tau_k - x} f_W(w)dw\right)$ *is differentiable w.r.t to* x *then it is Lipschitz continuous over* $|x| \leq x_{max}$ *with Lipschitz constant* $L_{g_k} = \sup_{|t|\leq x_{max}} \left|\frac{f_W(\tau_k - t) - f_W(\tau_{k-1}-t)}{\int_{\tau_{k-1}-t}^{\tau_k - t} f_W(w)dw}\right|$*, i.e., we have*

$$\Big|g_k(x) - g_k(y)\Big| \leq L_{g_k}|x-y|, \quad \forall |x|, |y| \leq x_{max}, \tag{4.24}$$

Proof. Assuming that the function $\log\left(\int_{\tau_{k-1}-x}^{\tau_k - x} f_W(w)dw\right)$ is differentiable w.r.t to x then we have

$$\begin{aligned}
\big|g_k(x) - g_k(y)\big| &= \left|\int_y^x \frac{dg_k(t)}{dt}dt\right| \\
&\leq \sup_{|t|\leq x_{max}} \left|\frac{dg_k(t)}{dt}\right| \left|\int_y^x dt\right| \\
&= \sup_{|t|\leq x_{max}} \left|\frac{dg_k(t)}{dt}\right| |x-y| \\
&= \sup_{|t|\leq x_{max}} \left|\frac{f_W(\tau_k - t) - f_W(\tau_{k-1}-t)}{\int_{\tau_{k-1}-t}^{\tau_k - t} f_W(w)dw}\right| |x-y|
\end{aligned}$$

□

Lemma 4.11.4. *We have*

$$\left|\log\frac{p(\mathbf{Z}_\mathcal{S};\mathbf{X}^*_\mathcal{S})}{p(\mathbf{Z}_\mathcal{S};\mathbf{Y}_\mathcal{S})} - \log\frac{p(\mathbf{Z}_\mathcal{S};\mathbf{X}^*_\mathcal{S})}{p(\mathbf{Z}_\mathcal{S};\mathbf{X}_\mathcal{S})}\right| \leq L_g|\mathcal{S}|\epsilon, \ \forall \mathbf{Y} \ \textit{satisfying} \ \|\mathbf{Y}-\mathbf{X}\|_\infty \leq \epsilon$$

$$\left|\mathbb{E}\log\frac{p(\mathbf{Z}_\mathcal{S};\mathbf{X}^*_\mathcal{S})}{p(\mathbf{Z}_\mathcal{S};\mathbf{Y}_\mathcal{S})} - \mathbb{E}\log\frac{p(\mathbf{Z}_\mathcal{S};\mathbf{X}^*_\mathcal{S})}{p(\mathbf{Z}_\mathcal{S};\mathbf{X}_\mathcal{S})}\right| \leq L_g\gamma n_1 n_2\epsilon, \ \forall \mathbf{Y} \ \textit{satisfying} \ \|\mathbf{Y}-\mathbf{X}\|_\infty \leq \epsilon$$

where $L_g := max_k \ L_{g_k}$.

Proof. 1. We have

$$\left|\log\frac{p(\mathbf{Z}_{\mathcal{S}};\mathbf{X}^*_{\mathcal{S}})}{p(\mathbf{Z}_{\mathcal{S}};\mathbf{Y}_{\mathcal{S}})}-\log\frac{p(\mathbf{Z}_{\mathcal{S}};\mathbf{X}^*_{\mathcal{S}})}{p(\mathbf{Z}_{\mathcal{S}};\mathbf{X})}\right|=\left|\log p(\mathbf{Z}_{\mathcal{S}};\mathbf{Y}_{\mathcal{S}})-\log p(\mathbf{Z}_{\mathcal{S}};\mathbf{X}_{\mathcal{S}})\right|$$

The log-likelihood term can be expanded as

$$\begin{aligned}\left|\log p(\mathbf{Z}_{\mathcal{S}};\mathbf{Y}_{\mathcal{S}})-\log p(\mathbf{Z}_{\mathcal{S}};\mathbf{X}_{\mathcal{S}})\right| &= \left|\sum_{(i,j)\in\mathcal{S}}\sum_{k=1}^{K}\mathbf{1}(Z_{i,j}=k)\log\left(\frac{p(Z_{i,j}=k;Y_{i,j})}{p(Z_{i,j}=k;X_{i,j})}\right)\right|\\ &= \left|\sum_{(i,j)\in\mathcal{S}}\sum_{k=1}^{K}\mathbf{1}(Z_{i,j}=k)\left(g_k(Y_{i,j})-g_k(X_{i,j})\right)\right|\\ &\leq \sum_{(i,j)\in\mathcal{S}}\sum_{k=1}^{K}\mathbf{1}(Z_{i,j}=k)\Big|\left(g_k(Y_{i,j})-g_k(X_{i,j})\right)\Big|\end{aligned}$$

Using lemma 4.11.3 we have

$$\left|\log p(\mathbf{Z}_{\mathcal{S}};\mathbf{Y}_{\mathcal{S}})-\log p(\mathbf{Z}_{\mathcal{S}};\mathbf{X}_{\mathcal{S}})\right|\leq\sum_{(i,j)\in\mathcal{S}}\sum_{k=1}^{K}\mathbf{1}(Z_{i,j}=k)L_{g_k}\left|Y_{i,j}-X_{i,j}\right|$$

Now for all $\|\mathbf{Y}-\mathbf{X}\|_\infty\leq\epsilon$ we have $\left|Y_{i,j}-X_{i,j}\right|\leq\epsilon$ for all (i,j) which implies

$$\begin{aligned}\left|\log p(\mathbf{Z}_{\mathcal{S}};\mathbf{Y}_{\mathcal{S}})-\log p(\mathbf{Z}_{\mathcal{S}};\mathbf{X}_{\mathcal{S}})\right| &\leq \sum_{(i,j)\in\mathcal{S}}\sum_{k=1}^{K}\mathbf{1}(Z_{i,j}=k)L_{g_k}\epsilon\\ &\leq (max_k\ L_{g_k})\sum_{(i,j)\in\mathcal{S}}\sum_{k=1}^{K}\mathbf{1}(Z_{i,j}=k)\epsilon\\ &= L_g|\mathcal{S}|\epsilon\end{aligned}$$

2. We have

$$\begin{aligned}\left|\mathbb{E}\log\frac{p(\mathbf{Z}_{\mathcal{S}};\mathbf{X}^*_{\mathcal{S}})}{p(\mathbf{Z}_{\mathcal{S}};\mathbf{Y}_{\mathcal{S}})}-\mathbb{E}\log\frac{p(\mathbf{Z}_{\mathcal{S}};\mathbf{X}^*_{\mathcal{S}})}{p(\mathbf{Z}_{\mathcal{S}};\mathbf{X}_{\mathcal{S}})}\right| &= \left|\mathbb{E}\log\frac{p(\mathbf{Z}_{\mathcal{S}};\mathbf{X}_{\mathcal{S}})}{p(\mathbf{Z}_{\mathcal{S}};\mathbf{Y}_{\mathcal{S}})}\right|\\ &\leq \mathbb{E}\left|\log\frac{p(\mathbf{Z}_{\mathcal{S}};\mathbf{X}_{\mathcal{S}})}{p(\mathbf{Z}_{\mathcal{S}};\mathbf{Y}_{\mathcal{S}})}\right|\\ &\leq L_g\epsilon\mathbb{E}|\mathcal{S}|=L_g\epsilon\gamma n_1n_2,\end{aligned}$$

where last step follows from second statement of the lemma.

□

Lemma 4.11.5. *Assuming $\gamma n_1 n_2 \geq 12\log(\frac{2}{\alpha})$, we have for the random variable $|\mathcal{S}| = \sum_{i,j} b_{ij}$*

$$Prob\left(\frac{1}{2}\gamma n_1 n_2 \leq |\mathcal{S}| \leq \frac{3}{2}\gamma n_1 n_2\right) \geq 1-\alpha$$

Proof. We first observe that $\mathbb{E}\left(|\mathcal{S}|\right) = \gamma n_1 n_2$ and then using relative Chernoff's bound we have $\text{Prob}\left(|\mathcal{S}| \geq \frac{3}{2}\gamma n_1 n_2\right) \leq e^{-\frac{\gamma n_1 n_2}{12}}$ and $\text{Prob}\left(|\mathcal{S}| \leq \frac{1}{2}\gamma n_1 n_2\right) \leq e^{-\frac{\gamma n_1 n_2}{8}}$. Rest of the proof follows the union bounding technique as follows

$$\begin{aligned}\text{Prob}\left(\frac{1}{2}\gamma n_1 n_2 \leq |\mathcal{S}| \leq \frac{3}{2}\gamma n_1 n_2\right) &= 1 - \text{Prob}\left(\left\{|\mathcal{S}| \leq \frac{1}{2}\gamma n_1 n_2\right\}\bigcup\left\{|\mathcal{S}| \geq \frac{3}{2}\gamma n_1 n_2\right\}\right)\\ &\geq 1 - \text{Prob}\left(|\mathcal{S}| \geq \frac{3}{2}\gamma n_1 n_2\right) - \text{Prob}\left(|\mathcal{S}| \leq \frac{1}{2}\gamma n_1 n_2\right)\\ &\geq 1 - e^{-\frac{\gamma n_1 n_2}{12}} - e^{-\frac{\gamma n_1 n_2}{8}}\end{aligned}$$

Further, if we choose $\gamma n_1 n_2 \geq 12\log\left(\frac{2}{\alpha}\right)$ then we have $e^{-\frac{\gamma n_1 n_2}{12}} \leq \alpha/2$ and $e^{-\frac{\gamma n_1 n_2}{8}} \leq \alpha/2$. This implies that if $\gamma n_1 n_2 \geq 12\log\left(\frac{2}{\alpha}\right)$ then we have

$$\text{Prob}\left(\frac{1}{2}\gamma n_1 n_2 \leq |\mathcal{S}| \leq \frac{3}{2}\gamma n_1 n_2\right) \geq 1-\alpha.$$

□

Lemma 4.11.6. *The covering number $N(\mathcal{X}, \|\cdot\|_\infty, \epsilon)$ for the set*

$$\mathcal{X} = \left\{\mathbf{X} = \mathbf{D}\mathbf{A}\middle|\mathbf{D} \in \mathcal{D}, \mathbf{A} \in \mathcal{A}, \|\mathbf{X}\|_\infty \leq x_{max}\right\}$$

where $\mathcal{D} \subseteq \left\{\mathbf{D} \in \mathbb{R}^{n_1 \times r}\middle|\|\mathbf{D}\|_\infty \leq d_{max}\right\}$ and $\mathcal{A} \subseteq \left\{\mathbf{A} \in \mathbb{R}^{r \times n_2}\middle|\|\mathbf{A}\|_\infty \leq a_{max}\right\}$ is upper bounded as

$$N(\mathcal{X}, \|\cdot\|_\infty, \epsilon) \leq \left(\frac{6a_{max}d_{max}r}{\epsilon}\right)^{n_1 r + n_2 r}.$$

Proof. Suppose we are given an ϵ_d cover $\mathcal{D}_{\epsilon_d}$ of the set $\mathcal{D}$ in $\|\cdot\|_\infty$ norm and an ϵ_a cover $\mathcal{A}_{\epsilon_a}$ of the set $\mathcal{A}$ in $\|\cdot\|_\infty$ norm. Consider any $\mathbf{X} \in \mathcal{X}$ which by construction can be factorized as $\mathbf{X} = \mathbf{D}_\mathbf{X}\mathbf{A}_\mathbf{X}$ where $\mathbf{D}_\mathbf{X} \in \mathcal{D}$ and $\mathbf{A}_\mathbf{X} \in \mathcal{A}$. Let $\mathbf{D}_{\epsilon_d}$ and $\mathbf{A}_{\epsilon_a}$ be the centers of the ball in which $\mathbf{D}_\mathbf{X}$ and $\mathbf{A}_\mathbf{X}$ lie, i.e., we have $\|\mathbf{D}_\mathbf{X} - \mathbf{D}_{\epsilon_d}\|_\infty \leq \epsilon_d$ and

$\|\mathbf{A_X} - \mathbf{A}_{\epsilon_a}\|_\infty \le \epsilon_a$. The main idea of the proof is find the ϵ of the covering provided by $\mathbf{D}_{\epsilon_d}$ and $\mathbf{A}_{\epsilon_a}$. For we consider

$$\begin{aligned}\|\mathbf{X} - \mathbf{D}_{\epsilon_d}\mathbf{A}_{\epsilon_a}\|_\infty &= \|\mathbf{D_X}\mathbf{A_X} - \mathbf{D}_{\epsilon_d}\mathbf{A}_{\epsilon_a}\|_\infty \\ &\le \frac{\|\mathbf{D_X}\left(\mathbf{A_X} - \mathbf{A}_{\epsilon_a}\right)\|_\infty + \|\left(\mathbf{D_X} - \mathbf{D}_{\epsilon_d}\right)\mathbf{A}_{\epsilon_a}\|_\infty}{2} \\ &\quad + \frac{\|\left(\mathbf{D_X} - \mathbf{D}_{\epsilon_d}\right)\mathbf{A_X}\|_\infty + \|\mathbf{D}_{\epsilon_d}\left(\mathbf{A_X} - \mathbf{A}_{\epsilon_a}\right)\|_\infty}{2}\end{aligned}$$

Next we explicitly write $\|\mathbf{D_X}\left(\mathbf{A_X} - \mathbf{A}_{\epsilon_a}\right)\|_\infty = \max_{i,j}\mathbf{d}_i^T(\mathbf{a}_j - \mathbf{a}_j^{\epsilon_a})$ where $\mathbf{d}_i^T$ is the i^{th} row of $\mathbf{D_X}$ and $\mathbf{a}_j, \mathbf{a}_j$ are the j^{th} column of matrices $\mathbf{A_X}$ and $\mathbf{A}_{\epsilon_a}$ respectively. Let $\mathbf{d}_i^{\epsilon_d T}$ be the i^{th} row $\mathbf{D}_{\epsilon_d}$. We continue from above as

$$\begin{aligned}\|\mathbf{X} - \mathbf{D}_{\epsilon_d}\mathbf{A}_{\epsilon_a}\|_\infty &\le \frac{\max_{i,j}\left|\mathbf{d}_i^T(\mathbf{a}_j - \mathbf{a}_j^{\epsilon_a})\right| + \max_{i,j}\left|(\mathbf{d}_i^{\epsilon_d} - \mathbf{d}_i)^T\mathbf{a}_j^{\epsilon_a}\right|}{2} \\ &\quad + \frac{\max_{i,j}\left|(\mathbf{d}_i^{\epsilon_d} - \mathbf{d}_i)^T\mathbf{a}_j\right| + \max_{i,j}\left|(\mathbf{d}_i^{\epsilon_d})^T(\mathbf{a}_j - \mathbf{a}_j^{\epsilon_a})\right|}{2} \\ &\le \frac{\max_i(\|\mathbf{d}_i\|_1 + \|\mathbf{d}_i^{\epsilon_d}\|_1)\epsilon_a + \max_i(\|\mathbf{a}_i\|_1 + \|\mathbf{a}_i^{\epsilon_a}\|_1)\epsilon_d}{2} \\ &\le r(d_{max}\epsilon_a + a_{max}\epsilon_d)\end{aligned}$$

The above in equality implies that $\mathcal{X}_{\epsilon_d\epsilon_a} = \left\{\mathbf{D}_{\epsilon_d}\mathbf{A}_{\epsilon_a}\middle|\mathbf{D}_{\epsilon_d} \in \mathcal{D}_{\epsilon_d}, \mathbf{A}_{\epsilon_a} \in \mathcal{A}_{\epsilon_a}\right\}$ provides covering for $\mathcal{X}$ with ϵ equal to $r(d_{max}\epsilon_a + a_{max}\epsilon_d)$. In other words, for a given ϵ covering for $\mathcal{X}$ we can use $\mathcal{X}_{\epsilon_d\epsilon_a}$ is as a cover such ϵ_a and ϵ_d satisfy

$$\epsilon = r(d_{max}\epsilon_a + a_{max}\epsilon_d)$$

With this we have the following bounds on the covering number $N(\mathcal{X}, \|\cdot\|_\infty, \epsilon)$

$$N(\mathcal{X}, \|\cdot\|_\infty, \epsilon) \le \left|\mathcal{X}_{\epsilon_d\epsilon_a}\right| = N(\mathcal{D}, \|\cdot\|_\infty, \epsilon_d)N(\mathcal{A}, \|\cdot\|_\infty, \epsilon_a)$$

Further choosing $\epsilon_a = \frac{\epsilon}{2rd_{max}}$ and $\epsilon_d = \frac{\epsilon}{2ra_{max}}$ we have $\epsilon = r(d_{max}\epsilon_a + a_{max}\epsilon_d)$ so we finally have

$$N(\mathcal{X}, \|\cdot\|_\infty, \epsilon) \le \left|\mathcal{X}_{\epsilon_d\epsilon_a}\right| = N\left(\mathcal{D}, \|\cdot\|_\infty, \frac{\epsilon}{2ra_{max}}\right)N\left(\mathcal{A}, \|\cdot\|_\infty, \frac{\epsilon}{2rd_{max}}\right)$$

Using standard arguments of covering numbers one can easily argue that

$$N\left(\mathcal{D}, \|\cdot\|_\infty, \frac{\epsilon}{2ra_{max}}\right) \leq \left(\frac{6a_{max}d_{max}r}{\epsilon}\right)^{n_1 r}$$
$$N\left(\mathcal{A}, \|\cdot\|_\infty, \frac{\epsilon}{2rd_{max}}\right) \leq \left(\frac{6a_{max}d_{max}r}{\epsilon}\right)^{n_2 r}$$

With the above two inequalities proof is complete.

□

Lemma 4.11.7. *The covering number $N(\mathcal{X}, \|\cdot\|_\infty, \epsilon)$ for the set*

$$\mathcal{X} = \left\{\mathbf{DA}\middle|\mathbf{D} \in \mathcal{D}, \mathbf{A} \in \mathcal{A}, \|\mathbf{X}\|_\infty \leq x_{max}\right\}$$

where

$$\mathcal{D} \subseteq \left\{\mathbf{D} \in \mathbb{R}^{n_1 \times r}\middle|\|\mathbf{D}\|_\infty \leq d_{max}\right\} \text{ and } \mathcal{A} \subseteq \left\{\mathbf{A} \in \mathbb{R}^{r \times n_2}\middle|\|\mathbf{A}\|_\infty \leq a_{max}, \|\mathbf{A}\|_0 \leq l\right\}$$

satisfies

$$N(\mathcal{X}, \|\cdot\|_\infty, \epsilon) \leq \left(\frac{n_2 re}{l}\right)^l \left(\frac{6a_{max}d_{max}r}{\epsilon}\right)^{n_1 r + l}.$$

Proof. Using the arguments presented in the proof of Lemma 4.11.6 we have

$$N(\mathcal{X}, \|\cdot\|_\infty, \epsilon) \leq N\left(\mathcal{D}, \|\cdot\|_\infty, \frac{\epsilon}{2ra_{max}}\right) N\left(\mathcal{A}, \|\cdot\|_\infty, \frac{\epsilon}{2rd_{max}}\right)$$

While the set $\mathcal{D}$ is same as in Lemma 4.11.6 but set $\mathcal{A}$ is different in this case. The proof is essentially about finding $N\left(\mathcal{A}, \|\cdot\|_\infty, \frac{\epsilon}{2rd_{max}}\right)$. For this we observe that $\mathcal{A} = \{\mathbf{A} = \mathbb{R}^{r \times n_2}\big|\|\mathbf{A}\|_\infty \leq a_{max}, \|\mathbf{A}\|_0 \leq l\} = \bigcup_{i=1}^{\binom{n_2 r}{l}} \mathcal{A}_i$, where each $\mathcal{A}_i$ corresponds to specific choice of support of l from $n_2 r$ possible location. Each $\mathcal{A}_i$ is a set of l dimensions with bounded entries as $\{\mathbf{a} \in \mathbb{R}^l\big|\|\mathbf{a}\|_\infty \leq a_{max}\}$. The covering number of $\mathcal{A}_i$ can be easily calculated as $N\left(\mathcal{A}_i, \|\cdot\|_\infty, \frac{\epsilon}{2rd_{max}}\right) \leq \left(\frac{6a_{max}d_{max}r}{\epsilon}\right)^l$. Since $\mathcal{A} = \bigcup_{i=1}^{\binom{n_2 r}{l}} \mathcal{A}_i$ we have

$$\begin{aligned} N\left(\mathcal{A}, \|\cdot\|_\infty, \frac{\epsilon}{2rd_{max}}\right) &\leq \sum_{i=1}^{\binom{n_2 r}{l}} N\left(\mathcal{A}_i, \|\cdot\|_\infty, \frac{\epsilon}{2rd_{max}}\right) \\ &\leq \binom{n_2 r}{l} \left(\frac{6a_{max}d_{max}r}{\epsilon}\right)^l \\ &\leq \left(\frac{n_2 re}{l}\right)^l \left(\frac{6a_{max}d_{max}r}{\epsilon}\right)^l \end{aligned}$$

The covering number of the set $\mathcal{D}$ is bounded as $N\left(\mathcal{D}, \|\cdot\|_\infty, \frac{\epsilon}{2rd_{max}}\right) \leq \left(\frac{6a_{max}d_{max}r}{\epsilon}\right)^{n_1 r}$. So finally we have

$$\begin{aligned}N(\mathcal{X}, \|\cdot\|_\infty, \epsilon) &\leq N\left(\mathcal{D}, \|\cdot\|_\infty, \frac{\epsilon}{2ra_{max}}\right) N\left(\mathcal{A}, \|\cdot\|_\infty, \frac{\epsilon}{2rd_{max}}\right)\\ &= \left(\frac{n_2 re}{l}\right)^l \left(\frac{6a_{max}d_{max}r}{\epsilon}\right)^{n_1 r + l}.\end{aligned}$$

□

4.11.2 Proof of lemma 4.5.1

Proof. We begin by observing that the random variable $U_{ij} = b_{ij} \log \frac{p(Z_{i,j}=k;X^*_{i,j})}{p(Z_{i,j}=k;X_{ij})}$ can be bounded from its expected value as follows

$$\begin{aligned}|U_{ij} - \mathbb{E}(U_{ij})| &\leq |U_{ij}| + |\mathbb{E}(U_{ij})|\\ &= b_{ij}\left|\sum_{k=1}^{K} \mathbf{1}(Z_{ij} = k) \log \frac{p(Z_{i,j} = k; X^*_{i,j})}{p(Z_{i,j} = k; X_{ij})}\right|\\ &+ \gamma\left|\sum_{k=1}^{K} p(Z_{i,j} = k; X^*_{i,j}) \log \frac{p(Z_{i,j} = k; X^*_{i,j})}{p(Z_{i,j} = k; X_{ij})}\right|\\ &\leq \log \frac{1-\delta_0}{\delta_0}\left(b_{ij}\left|\sum_{k=1}^{K} \mathbf{1}(Z_{ij} = k)\right| + \gamma\left|\sum_{k=1}^{K} p(Z_{i,j} = k; X^*_{i,j})\right|\right)\\ &\leq 2\log\Delta\end{aligned}$$

where $\Delta = \frac{1-\delta_0}{\delta_0}$. Equipped with boundedness of $|U_{ij} - \mathbb{E}(U_{ij})|$ we have from Craig Bernstein inequality [116]

$$\text{Prob}\left(\sum_{i,j} \mathbb{E}(U_{ij}) - U_{ij} \geq \frac{\tau}{c} + c\frac{\sum_{i,j} var(U_{ij})}{2(1-\theta)}\right) \leq e^{-\tau},$$

where $0 < \frac{2c\log\Delta}{3} \leq \theta < 1$. Using Lemma 4.11.2 we have $var(U_{ij}) \leq \omega\mathbb{E}(U_{ij})$ where $\omega = 2\Delta$. This gives us

$$\text{Prob}\left(\left[1 - \frac{c\omega}{2(1-\theta)}\right]\sum_{i,j} \mathbb{E}(U_{ij}) - U_{ij} \geq \frac{\tau}{c}\right) \leq e^{-\tau}.$$

We choose $\theta = \frac{1}{2}$ and $c = \frac{1}{2\omega}$ which reduces the condition $0 < \frac{2c\log\Delta}{3} \leq \theta < 1$ to $\frac{\log\Delta}{3\Delta} \leq 1$. Since under the model assumption $\delta_0 < \frac{1}{2}$ which insures that $\Delta > 1$ so the inequality $\frac{\log\Delta}{3\Delta} \leq 1$ is always satisfied. So finally we have

$$\text{Prob}\left(\frac{1}{2}\sum_{i,j}\mathbb{E}(U_{ij}) - U_{ij} \geq \frac{\tau}{c}\right) \leq e^{-\tau},$$

where $c = \frac{1}{2\omega}$.

□

4.11.3 Proof of lemma 4.5.2

Proof. We begin with by bounding the probability that the random variable $\log\frac{p(\mathbf{Z}_{\mathcal{S}};\mathbf{X}^*_{\mathcal{S}})}{p(\mathbf{Z}_{\mathcal{S}};\mathbf{X}_{\mathcal{S}})}$ is bounded away from its mean $\mathbb{E}\log\frac{p(\mathbf{Z}_{\mathcal{S}};\mathbf{X}^*_{\mathcal{S}})}{p(\mathbf{Z}_{\mathcal{S}};\mathbf{X}_{\mathcal{S}})}$ for all $\mathbf{X} \in \mathcal{X}_\epsilon$. Let $U_{ij} = b_{ij}\log\frac{p(Z_{i,j};X^*_{ij})}{p(Z_{i,j};X_{ij})}$ we can write as $\log\frac{p(\mathbf{Z}_{\mathcal{S}};\mathbf{X}^*_{\mathcal{S}})}{p(\mathbf{Z}_{\mathcal{S}};\mathbf{X}_{\mathcal{S}})} = \sum_{i,j} U_{i,j}$ we find the probability of following event

$$\text{Prob}\left(\sup_{\mathbf{X}\in\mathcal{X}_\epsilon}\frac{1}{2}\sum_{i,j}\mathbb{E}(U_{ij}) - \sum_{i,j}U_{ij} \leq \frac{\tau}{c}\right)$$

where c is as defined in the Lemma statement. We can write this probability as

$$\begin{aligned}
&\text{Prob}\left(\sup_{\mathbf{X}\in\mathcal{X}_\epsilon}\frac{1}{2}\sum_{i,j}\mathbb{E}(U_{ij}) - \sum_{i,j}U_{ij} \leq \frac{\tau}{c}\right)\\
&\quad= \text{Prob}\left(\bigcap_{\mathbf{X}\in\mathcal{X}_\epsilon}\frac{1}{2}\sum_{i,j}\mathbb{E}(U_{ij}) - \sum_{i,j}U_{ij} \leq \frac{\tau}{c}\right)\\
&\quad= 1 - \text{Prob}\left(\bigcup_{\mathbf{X}\in\mathcal{X}_\epsilon}\frac{1}{2}\sum_{i,j}\mathbb{E}(U_{ij}) - \sum_{i,j}U_{ij} \geq \frac{\tau}{c}\right)\\
&\quad\geq 1 - N(\epsilon, \|\cdot\|_\infty, \mathcal{X})\text{Prob}\left(\frac{1}{2}\sum_{i,j}\mathbb{E}(U_{ij}) - \sum_{i,j}U_{ij} \geq \frac{\tau}{c}\right).\\
&\quad\geq 1 - N(\epsilon, \|\cdot\|_\infty, \mathcal{X})e^{-\tau}
\end{aligned}$$

In above the second last step is due to union bound and the last step is by using lemma 4.5.1. Now we let $\alpha := N(\epsilon, \|\cdot\|_\infty, \mathcal{X})e^{-\tau}$ we can solve $\tau = \log \frac{N(\epsilon,\|\cdot\|_\infty,\mathcal{X})}{\alpha}$. So we have

$$\text{Prob}\left(\sup_{\mathbf{X}\in\mathcal{X}_\epsilon} \frac{1}{2}\sum_{i,j}\mathbb{E}(U_{ij}) - \sum_{i,j} U_{ij} \leq \frac{\log \frac{N(\epsilon,\|\cdot\|_\infty,\mathcal{X})}{\alpha}}{c}\right) \geq 1-\alpha \tag{4.25}$$

In other words, that with probability atleast $1-\alpha$ we have all $\mathbf{X} \in \mathcal{X}_\epsilon$

$$\frac{1}{2}\mathbb{E}\log\frac{p(\mathbf{Z}_\mathcal{S};\mathbf{X}^*_\mathcal{S})}{p(\mathbf{Z}_\mathcal{S};\mathbf{X}_\mathcal{S})} - \log\frac{p(\mathbf{Z}_\mathcal{S};\mathbf{X}^*_\mathcal{S})}{p(\mathbf{Z}_\mathcal{S};\mathbf{X}_\mathcal{S})} \leq \frac{\log \frac{N(\epsilon,\|\cdot\|_\infty,\mathcal{X})}{\alpha}}{c}$$

Next we extend the above probabilistic argument to full set $\mathcal{X}$ by using Lemma 4.11.4 which essentially bounds the value of $\log\frac{p(\mathbf{Z}_\mathcal{S};\mathbf{X}^*_\mathcal{S})}{p(\mathbf{Z}_\mathcal{S};\mathbf{X}_\mathcal{S})}$ and $\mathbb{E}\log\frac{p(\mathbf{Z}_\mathcal{S};\mathbf{X}^*_\mathcal{S})}{p(\mathbf{Z}_\mathcal{S};\mathbf{X}_\mathcal{S})}$ evaluated at points in a ball $B_\epsilon(\mathbf{X})$ associated with a center $\mathbf{X} \in \mathcal{X}_\epsilon$. Using Lemma 4.11.4 we have for $\mathbf{Y} \in B_\epsilon(\mathbf{X})$

$$\log\frac{p(\mathbf{Z}_\mathcal{S};\mathbf{X}^*_\mathcal{S})}{p(\mathbf{Z}_\mathcal{S};\mathbf{X}_\mathcal{S})} \leq \log\frac{p(\mathbf{Z}_\mathcal{S};\mathbf{X}^*_\mathcal{S})}{p(\mathbf{Z}_\mathcal{S};\mathbf{Y}_\mathcal{S})} + L_g|\mathcal{S}|\epsilon$$

$$\mathbb{E}\log\frac{p(\mathbf{Z}_\mathcal{S};\mathbf{X}^*_\mathcal{S})}{p(\mathbf{Z}_\mathcal{S};\mathbf{Y}_\mathcal{S})} - L_g\gamma n_1 n_2\epsilon \leq \mathbb{E}\log\frac{p(\mathbf{Z}_\mathcal{S};\mathbf{X}^*_\mathcal{S})}{p(\mathbf{Z}_\mathcal{S};\mathbf{X}_\mathcal{S})}$$

And since $\mathcal{X} = \bigcup_{\mathbf{X}_i\in\mathcal{X}_\epsilon} B_\epsilon(\mathbf{X}_i)$ we can say using the above two inequalities that for all $\mathbf{Y} \in \mathcal{X}$ with probability at least $1-\alpha$ we have

$$\frac{1}{2}\mathbb{E}\log\frac{p(\mathbf{Z}_\mathcal{S};\mathbf{X}^*_\mathcal{S})}{p(\mathbf{Z}_\mathcal{S};\mathbf{Y}_\mathcal{S})} - \log\frac{p(\mathbf{Z}_\mathcal{S};\mathbf{X}^*_\mathcal{S})}{p(\mathbf{Z}_\mathcal{S};\mathbf{Y}_\mathcal{S})} \leq \left(|\mathcal{S}| + \frac{\gamma n_1 n_2}{2}\right)L_g\epsilon + \frac{\log \frac{N(\epsilon,\|\cdot\|_\infty,\mathcal{X})}{\alpha}}{c}$$

□

4.12 Proof of lemma 4.6.1

Proof. We have

$$
\begin{aligned}
&|V_{\mathbf{X}} - V_{\mathbf{Y}}| \\
&= \left| \sum_{i,j} b_{ij} \sum_{k=1}^{K} \mathbf{1}(Z_{ij} = k) \log \frac{p(Z_{i,j} = k; Y_{i,j})}{p(Z_{i,j} = k; X_{ij})} - \gamma \sum_{i,j} \sum_{k=1}^{K} p(Z_{ij} = k; X^*_{ij}) \log \frac{p(Z_{i,j} = k; Y_{i,j})}{p(Z_{i,j} = k; X_{ij})} \right| \\
&= \left| \sum_{i,j} \sum_{k=1}^{K} \big(b_{ij} \mathbf{1}(Z_{ij} = k) - \gamma p(Z_{ij} = k; X^*_{ij}) \big) \log \frac{p(Z_{i,j} = k; Y_{i,j})}{p(Z_{i,j} = k; X_{ij})} \right| \\
&\le \sum_{i,j} \sum_{k=1}^{K} \left| \big(b_{ij} \mathbf{1}(Z_{ij} = k) - \gamma p(Z_{ij} = k; X^*_{ij}) \big) \right| L_{g_k} \left| Y_{ij} - X_{ij} \right| \\
&\le \|\mathbf{Y} - \mathbf{X}\|_F \|\mathbf{C}\|_F, \quad \text{where } C_{ij} = \sum_{k=1}^{K} \big(b_{ij} \mathbf{1}(Z_{ij} = k) - \gamma p(Z_{ij} = k; X^*_{ij}) \big) L_{g_k} \\
&\le \|\mathbf{Y} - \mathbf{X}\|_F \sqrt{n_1 n_2} L_g (1 + \gamma),
\end{aligned}
$$

where in last step is due to $|C_{ij}| \le L_g(1+\gamma)$ where $L_g = \max_k L_{g_k}$. The above inequality implies that the increments $|V_{\mathbf{X}} - V_{\mathbf{Y}}|$ are sub-gaussian with $\sigma = L_g \sqrt{n_1 n_2}\,(\gamma + 1)\, \|\mathbf{Y} - \mathbf{X}\|_F$, i.e., we have

$$\text{Prob}\left(|V_{\mathbf{X}} - V_{\mathbf{Y}}| \ge u \right) \le 2 e^{-\frac{2u^2}{L_g^2 n_1 n_2 (1+\gamma)^2 \|\mathbf{X}-\mathbf{Y}\|_F^2}}.$$

Further, with the change of variable $\tau^2 = \frac{2u^2}{L_g^2 (1+\gamma)^2 n_1 n_2 \|\mathbf{X}-\mathbf{Y}\|_F^2}$ and using the fact that $1 + \gamma \le 2$ we have

$$\text{Prob}\left(|V_{\mathbf{X}} - V_{\mathbf{Y}}| \ge \tau L_g \sqrt{2 n_1 n_2} \|\mathbf{X} - \mathbf{Y}\|_F \right) \le 2 e^{-\tau^2}.$$

□

4.13 Proof of lemma 4.6.2

Proof. Using the Theorem 3.2 from [117] we have if $U_{\mathbf{X}}$ satisfies

$$\text{Prob}\left(|V_{\mathbf{X}} - V_{\mathbf{Y}}| \ge u d(\mathbf{X}, \mathbf{Y}) \right) \le 2 e^{-u^2}$$

then for any $\mathbf{X}_0 \in \mathcal{X}$

$$\text{Prob}\left(\sup_{\mathbf{X} \in \mathcal{X}} |V_{\mathbf{X}} - V_{\mathbf{X}_0}| \ge \sqrt{e}\, \big(C \gamma_2(T, d) + u D \Delta_d(T) \big) \right) \le e^{-u^2/2}, \tag{4.26}$$

where

$$\Delta_d(\mathcal{X}) = \sup_{\mathbf{X},\mathbf{Y}\in\mathcal{X}} d(\mathbf{X},\mathbf{Y})$$
$$\gamma_2(\mathcal{X}, d) = \inf_{\mathcal{X}_a} \sup_{\mathbf{X}\in\mathcal{X}} \sum_{n=0}^{\infty} 2^{n/2} \inf_{\mathbf{Z}\in X_n} d(\mathbf{X},\mathbf{Z})$$

and $\mathcal{X}_a = (\mathcal{X}_n)_{n\geq 0}$ is a sequence of subset of $\mathcal{X}$ which satisfies $|\mathcal{X}_0| = 1$ and $|\mathcal{X}_n| \leq 2^{2^n}$ for all $n \geq 1$. We use this result for the random process indexed by $\mathbf{X} \in \mathcal{X}$ defined as follows

$$V_{\mathbf{X}} = \sum_{i,j} b_{ij} \log \frac{p(Z_{i,j}; X^*_{i,j})}{p(Z_{i,j}; X_{ij})} - \mathbb{E}\left(\sum_{i,j} b_{ij} \log \frac{p(Z_{i,j}; X^*_{i,j})}{p(Z_{i,j}; X_{ij})}\right).$$

For the above random process using the inequality in Lemma 4.6.1 we can invoke the above result with $d(\mathbf{X},\mathbf{Y}) = \frac{L_g(1+\gamma)\sqrt{n_1 n_2}\|\mathbf{X}-\mathbf{Y}\|_F}{\sqrt{2}}$ and $\gamma_2(\mathbf{X}, d), \Delta_d(\mathcal{X})$ can be simplified as follows

$$\gamma_2(\mathcal{X}, d) = L_g\sqrt{2n_1 n_2} \underbrace{\inf_{\mathcal{X}_a} \sup_{\mathbf{X}\in\mathcal{X}} \sum_{n=0}^{\infty} 2^{n/2} \inf_{Z\in X_n} \|\mathbf{Z}-\mathbf{X}\|_F}_{:=\gamma_2(\mathcal{X},\ell_2)},$$
$$\Delta_d(\mathcal{X}) = L_g\sqrt{2n_1 n_2} \underbrace{\sup_{\mathbf{X},\mathbf{Y}\in\mathcal{X}} \|\mathbf{X}-\mathbf{Y}\|_F}_{:=\Delta_{\ell_2}(\mathcal{X})}$$

Now using the concentration inequality in (4.26) we have

$$\text{Prob}\left(\sup_{\mathbf{X}\in\mathcal{X}} |V_{\mathbf{X}} - V_{\mathbf{X}_0}| \geq \sqrt{e} L_g \sqrt{2n_1 n_2}\left(C\gamma_2(\mathcal{X},\ell_2) + uD\Delta_{\ell_2}(\mathcal{X})\right)\right) \leq e^{-u^2/2},$$

Further, redefining $C := C\sqrt{2e}, D := D\sqrt{2e}$ and substituting $\mathbf{X}_0 = \mathbf{X}^*$ for which $U_{\mathbf{X}_0} = 0$, we have

$$\text{Prob}\left(\sup_{\mathbf{X}\in\mathcal{X}} |V_{\mathbf{X}}| \geq L_g \sqrt{2n_1 n_2}\left(C\gamma_2(\mathcal{X},\ell_2) + uD\Delta_{\ell_2}(\mathcal{X})\right)\right) \leq e^{-u^2/2}.$$

□

4.14 Proof of corollary 4.6.1

Proof. Using the definition of $\mathcal{X}$ in (4.14) the Gaussian width $\Omega(\mathcal{X})$ can be calculated as follows

$$\begin{aligned}
\Omega(\mathcal{X}) &= \mathbb{E}\left[\sup_{\mathbf{a}_i\in\mathbb{R}^{n_2},\|\mathbf{a}_i\|_2\leq 1,\mathbf{d}_i\in\mathbb{R}^{n_1},\|\mathbf{d}_i\|_2\leq 1} \operatorname{Tr}\left(\mathbf{G}\sum_{i=1}^{r}\mathbf{d}_i\mathbf{a}_i^T\right)\right] \\
&= r\mathbb{E}\left[\sup_{\mathbf{a}\in\mathbb{R}^{n_2},\|\mathbf{a}\|_2\leq 1,\mathbf{d}\in\mathbb{R}^{n_1},\|\mathbf{d}\|_2\leq 1} \operatorname{Tr}\left(\mathbf{G}\mathbf{d}\mathbf{a}^T\right)\right] \\
&= r\mathbb{E}\left[\|\mathbf{G}\|_2\right] = r(\sqrt{n_1}+\sqrt{n_2}) \qquad (4.27)
\end{aligned}$$

where the last step is due theorem 5.32 in [118]. Further, the $\Delta(\mathcal{X},\ell_2)$ can be bounded as follows

$$\begin{aligned}
\Delta(\mathcal{X},\ell_2) &\leq 2\sup_{\mathbf{X}\in\mathcal{X}}\|\mathbf{X}\|_F \\
&= 2r\left[\sup_{\mathbf{a}\in\mathbb{R}^{n_2},\|\mathbf{a}\|_2\leq 1,\mathbf{d}\in\mathbb{R}^{n_1},\|\mathbf{d}\|_2\leq 1}\|\mathbf{d}\mathbf{a}^T\|_F\right] \\
&\leq 2r\left[\sup_{\mathbf{a}\in\mathbb{R}^{n_2},\|\mathbf{a}\|_2\leq 1,\mathbf{d}\in\mathbb{R}^{n_1},\|\mathbf{d}\|_2\leq 1}\|\mathbf{d}\|_1\|\mathbf{a}\|_2\right],
\end{aligned}$$

where is last step is due the triangle inequality. Further, using the fact that $\|\mathbf{d}\|_1 \leq \sqrt{n_1}\|\mathbf{d}\|_2$ we have

$$\Delta(\mathcal{X},\ell_2) \leq 2r\sqrt{n_1}. \qquad (4.28)$$

Using the upper bounds on $\Omega(\mathcal{X})$ and $\Delta(\mathcal{X},\ell_2)$ the bound in theorem 4.6.1 we have the following performance bound

$$\begin{aligned}
\frac{\mathrm{D}(p(\mathbf{Z};\mathbf{X}^*)||p(\mathbf{Z};\hat{\mathbf{X}}))}{n_1n_2} &\leq \frac{L_g\left[C\beta r(\sqrt{n_1}+\sqrt{n_2})+\sqrt{2\log\left(\frac{1}{\alpha}\right)}D2r\sqrt{n_1}\right]}{\gamma\sqrt{n_1n_2}} \\
&\leq \frac{L_g\left[C\beta+\sqrt{2\log\left(\frac{1}{\alpha}\right)}D\right]r(\sqrt{n_1}+\sqrt{n_2})}{\gamma\sqrt{n_1n_2}} \\
&\leq \frac{L_g\left[C\beta+\sqrt{2\log\left(\frac{1}{\alpha}\right)}D\right]r(n_1+n_2)}{\gamma\sqrt{n_1n_2}}.
\end{aligned}$$

□

Chapter 5

Compressive measurement designs for estimating structured signals in structured clutter

In statistical estimation, tasks arising, for example, in compressive sensing (CS), we are often equipped with prior knowledge about the object we wish to infer (e.g., smoothness, characterized by the presence of only low-frequency components in the Fourier domain; a priori region of interest knowledge; or shared features extracted from sets of training data similar to the signal being acquired). [1] This work considers an estimation task in compressive sensing, where the goal is to estimate an unknown signal from compressive measurements that are corrupted by additive pre-measurement noise (interference, or clutter) as well as post-measurement noise, in the specific setting where some prior knowledge on the signal, interference, and noise is available. Let $\mathbf{x} \in \mathbb{R}^n$ represent the object we aim to estimate, and suppose that we obtain m noisy measurements of $\mathbf{x}$ as follows

$$\mathbf{y} = \mathbf{A}(\mathbf{x} + \mathbf{c}) + \mathbf{w}, \tag{5.1}$$

where $\mathbf{A}$ is the $m \times n$ sensing matrix, $\mathbf{c}$ as a $n \times 1$ vector of pre-measurement

[1] The material in this chapter is ©2013 IEEE. Reprinted, with permission, from *Asilomar Conference on Signals, Systems and Computers*, "Compressive measurement designs for estimating structured signals in structured clutter: A Bayesian Experimental Design approach", S. Jain, A. Soni, and J. Haupt.

interference or "clutter," and $\mathbf{w} \in \mathbb{R}^m$ is a vector of perturbations whose elements may describe additive measurement noise or modeling error.

Investigations of problems of this form in the so-called under-determined setting ($m < n$) have been the primary focus of recent efforts in CS research such as the analysis of sensing and inference procedures for estimating $\mathbf{x}$ from noisy linear measurements in case where $\mathbf{x}$ is sparse, having, say, $k < n$ nonzero or significant entries.

Investigations of problems of this form in the so-called under-determined setting ($m < n$) have been the primary focus of recent efforts in CS research in which a primary focus has been on the analysis of sensing and inference procedures for estimating $\mathbf{x}$ from such noisy linear measurements in the case where $\mathbf{x}$ is *sparse*, having, say, $k < n$ nonzero or significant entries. A notable aspect of the result [119] (in the clutter-free scenario) and indeed, many related results in the CS literature, is that of "universality" in which the *random* matrices $\mathbf{A}$ whose elements are drawn iid from certain zero-mean distributions comprise a broad class of sensing matrices that facilitate accurate estimation of sparse $\mathbf{x}$ in CS (see, for example, [120]).

On the other hand, in many scenarios we may be equipped with additional information about the signal we aim to estimate, beyond simply an assumption of sparsity. This additional information can be incorporated into the inference task to improve estimation performance [121,122]. A unique (in fact, *essential*) assumption underlying the CS paradigm is the ability to obtain *general* linear measurements of the quantity of interest. This inherent flexibility of the measurement process suggests that we should consider incorporating (in a principled manner, and as appropriate) the additional information directly into the design of the *sensing* process.

Motivated by this here we focus on a *knowledge-enhanced* estimation problem associated with the compressive measurements obtained via the model (5.1). The question we address here is, how should we design the sensing matrix $\mathbf{A}$ to take advantage of this prior knowledge?

5.1 Problem statement and our contributions

As alluded above our ultimate inference goal is to accurately estimate the vector $\mathbf{x}$ given measurements obtained according to (5.1), in settings where we may design the sensing matrix $\mathbf{A}$ using prior information about the signal, clutter, and noise. We focus on designing the sensing matrix by minimizing the mean-square error (MSE) associated with the ultimate estimate of the signal $\mathbf{x}$. Formally, we denote by $\widehat{\mathbf{x}}_{\mathbf{A}}(\mathbf{y})$ an *estimate* of $\mathbf{x}$ obtained using a particular estimation *strategy*, denoted here by $\widehat{\mathbf{x}}_{\mathbf{A}}$. Note that the estimation strategy is parameterized by the sensing matrix $\mathbf{A}$, and a particular estimate obtained using this strategy is a function of the measurements $\mathbf{y}$ obtained via (5.1) using that $\mathbf{A}$. The mean-square error associated with a particular estimation strategy $\widehat{\mathbf{x}}_{\mathbf{A}}$ is $\mathbb{E}_{\mathbf{x},\mathbf{c},\mathbf{w}}\left[\|\mathbf{x}-\widehat{\mathbf{x}}_{\mathbf{A}}(\mathbf{y})\|^2\right]$, where the subscript denotes that the expectation is with respect to all of the random quantities. The criteria for optimal design of the sensing matrix $\mathbf{A}$ in this case can be stated as an optimization – the optimal choice of $\mathbf{A}$, denoted by $\mathbf{A}^*$, is

$$\mathbf{A}^* = \arg\min_{\mathbf{A}\in\mathcal{A}} \min_{\widehat{\mathbf{x}}_{\mathbf{A}}\in\mathcal{X}} \mathbb{E}_{\mathbf{x},\mathbf{c},\mathbf{w}}\left[\|\mathbf{x}-\widehat{\mathbf{x}}_{\mathbf{A}}(\mathbf{y})\|^2\right], \tag{5.2}$$

where $\mathcal{A}$ is a (possibly constrained) class of sensing matrices and $\mathcal{X}$ is a (possibly constrained) class of possible estimation strategies. In words, $\mathbf{A}^* \in \mathcal{A}$ is the sensing matrix yielding measurements for which the MSE of the *best possible* estimation strategy (from the class $\mathcal{X}$) is minimum.

The problem as stated in (5.2) is very general. Depending on the application at hand it may vary depending on the sets $\mathbf{A}$, $\mathbf{X}$ and the prior information available at hand. Our works [5] and [6] investigated this problem for specific cases where the prior knowledge about the signal $\mathbf{x}$, clutter $\mathbf{c}$ and noise $\mathbf{w}$ is restricted up-to first- and second-order statistics, the class of estimator $\mathcal{X}$ was restricted to linear estimators, and the set $\mathcal{A}$ was the set of matrices with sensing energy constraints. Our initial work [5] addressed the sensing matrix design problem when noise power is high which allowed for certain simplification optimization problem. A general solution without any assumption on noise power was proposed in [6]. Due to the space limitations here we only report the details of [6].

5.2 Connections with existing work

In [123], the authors propose one of the first approaches to design compressive sensing matrices given some prior signal knowledge. The authors considered noise-free settings and assumed knowledge of a dictionary in which the signals being observed were sparse, and proposed a sensing matrix design procedure whose aim is to reduce the coherence between the learned sensing matrix and the known dictionary. Extensions of this idea aimed at designing both the dictionary and the sensing matrix given a collection of training data were examined by [124,125]. The work [126] studied knowledge-enhanced CS design tasks using a probabilistic formulation of the prior knowledge under the assumption that signal has Gaussian mixture prior, and proposed a design criteria based on coherence minimization between the learned sensing matrix and a dictionary composed of eigenvectors of the mixture covariance matrices. Along the same lines, the work [127] examined sensing designs based on learned correlations in training data. We note that none of these approaches utilize the statistical estimation theoretic formulation we adopt here.

Several recent works have examined the effects of clutter (i.e., the case $\mathbf{c} \neq \mathbf{0}$) in compressive sensing estimation tasks, but these investigations have typically been limited to the case where the clutter is modeled as Gaussian noise [128–131]. Several related work are along Bayesian experimental design problems in compressive sensing estimation tasks [132–136], and subsequent efforts [137,138] along these lines examined the performance improvements resulting from Bayesian experimental design strategies. Our problem is also related to the wealth of classical work on *interference cancellation* (see, for example, [139]). Prior work on optimal designs for space-time linear coding in MIMO applications – see, for example, [140], have examined qualitatively similar estimation problems but without the additive interference or “clutter” term.

5.3 Quantifying prior information

We assume that the signal $\mathbf{x} \in \mathbb{R}^n$ is a random quantity drawn from a mixture distribution having m_x mixture components. We do not assume full knowledge of the mixture distribution, but only that i-th mixture component has known weight $\pi_{x,i}$ and is an n-dimensional zero-mean random vector with known $n \times n$ covariance matrix $\mathbf{\Sigma}_{x,i}$, for $i = 1, 2, \ldots, m_x$. We note that the covariance matrices $\mathbf{\Sigma}_{x,i}$ are not assumed here to be full-rank. On the contrary, rank-deficiency in any of the $\mathbf{\Sigma}_{x,i}$ amounts to a form of *sparsity*, as random vectors $\mathbf{x} \in \mathbb{R}^n$ drawn from a distribution with covariance matrix of rank $r < n$ inherently lie on a r-dimensional subspace of $\mathbb{R}^n$. Thus, the formulation described here can model various forms of sparsity and structure that have been studied in the literature, block sparsity, group sparsity (with potentially overlapping groups), tree sparsity, and so on. Likewise, we assign an analogous prior distribution to the clutter $\mathbf{c}$, modeling it as a realization of an m_c-component mixture distribution whose i-th mixture component has weight $\pi_{c,i}$ and is a zero-mean random vector with covariance matrix $\mathbf{\Sigma}_{c,i}$, for $i = 1, 2, \ldots, m_c$. We consider $\mathbf{w}$ to be additive uncorrelated zero-mean noises with unit variance, and we assume that the random quantities $\mathbf{x}$, $\mathbf{c}$, and $\mathbf{w}$ are uncorrelated.

5.4 Choosing the set of sensing matrices $\mathcal{A}$

The presence of the measurement noise $\mathbf{w}$ is only relevant when the sensing matrix $\mathbf{A}$ is constrained in some way. Indeed, in unconstrained settings simply scaling each of the elements of $\mathbf{A}$ toward infinity would make the overall effect on $\mathbf{w}$ negligible in the estimation task. Here our focus will be on *sensing energy-constrained* designs $\mathbf{A}$; in particular, we choose $\mathcal{A}$ in (5.2) as $\mathcal{A} = \{\mathbf{A} : \|\mathbf{A}\|_F \leq \alpha\}$ for some (specified) $\alpha > 0$, where the notation $\|\cdot\|_F$ denotes the matrix Frobenius norm. Each row of $\mathbf{A}$ is itself a linear operator which gives rise to one (noisy, cluttered) compressive sample; thus, the constraint we impose here amounts to a constraint on the average energy per-row in the sensing matrix.

5.5 Choosing set of estimators $\mathcal{X}$

It is well-known from statistical estimation theory that, for the minimum MSE task (MMSE) task described above, the optimal estimator of $\mathbf{x}$ is the conditional mean $\mathbf{x}$ given the observations $\mathbf{y}$; that is, $\widehat{\mathbf{x}}_{\mathbf{A},\mathrm{MMSE}}(\mathbf{y}) = \mathbb{E}\left[\mathbf{x}|\mathbf{y}\right]$ (see, for example, [141]). Here, our prior knowledge is limited to first- and second-order statistics of the signal, clutter, and noise, and without full knowledge of the distributions we are unable to compute this estimator in closed form. Instead, we consider restricting the class of estimators $\mathcal{X}$ in (5.2) to be the class of *linear* estimators of $\mathbf{x}$.

We define the *average* signal covariance matrix $\mathbf{\Sigma}_x$ as $\mathbf{\Sigma}_x = \sum_{i=1}^{m_x} \pi_{x,i}\mathbf{\Sigma}_{x,i}$, and similarly for $\mathbf{\Sigma}_c$, and we assume that $(\mathbf{\Sigma}_x + \mathbf{\Sigma}_c)$ is invertible. Now, the linear MMSE estimator is just the Wiener filter, easily shown here to be

$$\widehat{\mathbf{x}}_{\mathbf{A},\text{LMMSE}}(\mathbf{y}) = \boldsymbol{\Sigma}_x \mathbf{A}' \left(\mathbf{A}\left(\boldsymbol{\Sigma}_x + \boldsymbol{\Sigma}_c\right)\mathbf{A}' + \mathbf{I}_n\right)^{-1} \mathbf{y},$$

where $\mathbf{A}'$ denotes the matrix transpose. It follows (after a bit of algebra) that

$$\mathbb{E}_{\mathbf{x},\mathbf{c},\mathbf{w}}\left[\|\mathbf{x} - \widehat{\mathbf{x}}_{\mathbf{A},\text{LMMSE}}(\mathbf{y})\|^2\right] = \text{tr}\{\boldsymbol{\Sigma}_{\text{x}} - \boldsymbol{\Sigma}_{\text{x}}\mathbf{A}'\left(\mathbf{A}\left(\boldsymbol{\Sigma}_{\text{x}} + \boldsymbol{\Sigma}_{\text{c}}\right)\mathbf{A}' + \mathbf{I}_{\text{m}}\right)^{-1}\mathbf{A}\boldsymbol{\Sigma}_{\text{x}}\}, \quad (5.3)$$

where tr$\{\cdot\}$ denotes the matrix trace (the sum of the diagonal elements) and $\mathbf{A}'$ is the transpose of $\mathbf{A}$. Thus, we can express our sensing matrix design task as an optimization, whose aim is to minimize the trace of the estimation error covariance matrix. In the parlance of Bayesian experimental design, this corresponds to a simple instance of a Bayes A-optimality criteria. Here, this amounts to an optimization problem

$$\mathbf{A}^* = \arg \max_{\mathbf{A}:\|\mathbf{A}\|_F \leq \alpha} \text{tr}\left\{\boldsymbol{\Sigma}_{\text{x}}\mathbf{A}'\left(\mathbf{A}\left(\boldsymbol{\Sigma}_{\text{x}} + \boldsymbol{\Sigma}_{\text{c}}\right)\mathbf{A}' + \mathbf{I}_{\text{m}}\right)^{-1}\mathbf{A}\boldsymbol{\Sigma}_{\text{x}}\right\}. \quad (5.4)$$

A similar problem was addressed in [142], but under a *transmit energy* constraint of the form tr $\left(\mathbf{A}\left(\boldsymbol{\Sigma}_{\text{x}} + \boldsymbol{\Sigma}_{\text{c}}\right)\mathbf{A}^{\text{T}}\right) \leq \alpha^2$. That said, the solution approach therein seems to be specific to the *transmit energy* constraint, and can not be directly extended to address the *sensing energy* constraint tr $\left(\mathbf{A}\mathbf{A}^{\text{T}}\right) \leq \alpha^2$ we impose here.

Our preliminary investigation on this problem (reported in [143]) entailed a solution approach for (5.4) that utilized an approximation of the inverse term in the objective, and led to a design strategy whose applicability was valid only in qualitatively low-SNR regimes. This was extended in [5] where a simple procedure for obtaining the solution to (5.4) was proposed, and performance improvements resulting from resulting approach was demonstrated via simulations. Next we describe the design approach proposed in [5].

5.6 Our proposed design approach

For solving (5.4) we make the following variable transformation: let $\mathbf{A}' = \mathbf{Y}\mathbf{M}$, where $\mathbf{Y}$ is $n \times n$ full rank matrix satisfying $\mathbf{Y}'(\boldsymbol{\Sigma}_x + \boldsymbol{\Sigma}_c)\mathbf{Y} = \mathbf{I}_n$, and $\mathbf{M}$ any $n \times m$ matrix. Since $\boldsymbol{\Sigma}_x + \boldsymbol{\Sigma}_c$ is positive definite, we can always find a diagonalizing matrix $\mathbf{Y}$ from the eigenvalue decomposition of $\boldsymbol{\Sigma}_x + \boldsymbol{\Sigma}_c$. Specifically, let $\boldsymbol{\Sigma}_x + \boldsymbol{\Sigma}_c = \mathbf{U}_{\mathbf{x+c}}\boldsymbol{\Sigma}_{\mathbf{x+c}}\mathbf{U}'_{\mathbf{x+c}}$, then $\mathbf{Y} = \mathbf{U}_{\mathbf{x+c}}\Sigma_{\mathbf{x+c}}^{-1/2}$. Overall there can be many choices of $\mathbf{Y}$ which diagonalize $\boldsymbol{\Sigma}_x + \boldsymbol{\Sigma}_c$;

in fact, $\mathbf{\Sigma}_x + \mathbf{\Sigma}_c$ is diagonalized by $\mathbf{Y}$ then it is also diagonalized by $\mathbf{YQ}$ for any orthonormal matrix $\mathbf{Q}$. Further, using the thin singular value decomposition of $\mathbf{M} = \mathbf{U}_M \mathbf{\Sigma}_M \mathbf{V}'_M$, where $\mathbf{U}_M \in \mathbb{R}^{n\times m}$ with $\mathbf{U}'_M \mathbf{U}_M = \mathbf{I}_m$, $\mathbf{\Sigma}_M = \mathbf{Diag}(\sigma_1, \cdots, \sigma_m)$ with $\sigma_i \geq 0 \quad \forall i = 1 \ldots m$, and $\mathbf{V}_M \in \mathbb{R}^{m\times m}$ is an orthonormal matrix, we can recast the problem (5.4) after a bit of linear algebra as

$$\begin{array}{ll} \underset{\substack{\mathbf{U}_M \in \mathbb{R}^{n\times m} \\ \sigma_i \geq 0}}{\text{maximize}} & \operatorname{tr}\left(\mathbf{U}'_{\mathrm{M}} \mathbf{Y}' \mathbf{\Sigma}^2_{\mathrm{x}} \mathbf{Y} \mathbf{U}_{\mathrm{M}} \tilde{\mathbf{\Sigma}}\right) \\ \text{subject to} & \mathbf{U}'_M \mathbf{U}_M = \mathbf{I}_m,\ \operatorname{tr}(\mathbf{U}'_{\mathrm{M}} \mathbf{Y}' \mathbf{Y} \mathbf{U}_{\mathrm{M}} \mathbf{\Sigma}^2_{\mathrm{M}}) \leq \alpha^2, \end{array} \tag{5.5}$$

where $\tilde{\mathbf{\Sigma}} = \mathbf{Diag}\left(\frac{\sigma_1^2}{1+\sigma_1^2}, \cdots, \frac{\sigma_m^2}{1+\sigma_m^2}\right)$. The problem (5.5) is a non-convex problem, so we resort to successive minimization over $\{\sigma_i\}_{i=1}^m$ and $\mathbf{U}_M$ by successively solving the following subproblems

$$\begin{array}{lll} \mathbf{P_0}: & \underset{\{\sigma_i\}_{i=1}^m;\ \sigma_i \geq 0\ \forall i}{\text{maximize}} & \operatorname{tr}\left(\mathbf{U}'_{\mathrm{M}} \mathbf{Y}' \mathbf{\Sigma}^2_{\mathrm{x}} \mathbf{Y} \mathbf{U}_{\mathrm{M}} \tilde{\mathbf{\Sigma}}\right) \\ & \text{subject to} & \operatorname{tr}(\mathbf{\Sigma}_{\mathrm{M}} \mathbf{U}'_{\mathrm{M}} \mathbf{Y}' \mathbf{Y} \mathbf{U}_{\mathrm{M}} \mathbf{\Sigma}_{\mathrm{M}}) \leq \alpha^2. \\ \mathbf{P_1}: & \underset{\mathbf{U}_M \in \mathbb{R}^{n\times m}}{\text{maximize}} & \operatorname{tr}\left(\mathbf{U}'_{\mathrm{M}} \mathbf{Y}' \mathbf{\Sigma}^2_{\mathrm{x}} \mathbf{Y} \mathbf{U}_{\mathrm{M}} \tilde{\mathbf{\Sigma}}\right) \\ & \text{subject to} & \mathbf{U}'_M \mathbf{U}_M = \mathbf{I}_m. \end{array} \tag{5.6}$$

```
Algorithm to solve: $\max_{\mathbf{A}\in\mathbb{R}^{m\times n}, \|\mathbf{A}\|_F^2\leq\alpha^2} \mathrm{tr}\left(\mathbf{\Sigma}_{\mathrm{x}}\mathbf{A}\left(\mathbf{A}\left(\mathbf{\Sigma}_{\mathrm{x}}+\mathbf{\Sigma}_{\mathrm{c}}\right)\mathbf{A}'+\mathbf{I}\right)^{-1}\mathbf{A}'\mathbf{\Sigma}_{\mathrm{x}}'\right)$
Input: Covariance matrices $\mathbf{\Sigma}_x, \mathbf{\Sigma}_c$, budget parameter $\alpha$, number of iterations $N$
1: Find $\mathbf{Y}$ such that $\mathbf{Y}'(\mathbf{\Sigma}_x+\mathbf{\Sigma}_c)\mathbf{Y}=\mathbf{I}_m$.
2: Calculate eigendecomposition of $\mathbf{Y}'\mathbf{\Sigma}_x^2\mathbf{Y}$, denoted $\mathbf{U}_1\mathbf{\Lambda}_1\mathbf{U}_1'=\mathbf{Y}'\mathbf{\Sigma}_x^2\mathbf{Y}$
3: Initialize $\{\sigma_i^0\}_{i=1}^m$
Repeat: 4 and 5 for $j=1$ to $N$
  4: Update $\mathbf{U}_M^j$ by $m$ columns of $\mathbf{U}_1$ by maximizing the following
         $\mathrm{tr}\left(\mathbf{U}_{\mathrm{M}}'\mathbf{Y}'\mathbf{\Sigma}_{\mathrm{x}}^2\mathbf{Y}\mathbf{U}_{\mathrm{M}}\mathbf{Diag}\left(\frac{(\sigma_1^{\mathrm{j}})^2}{1+(\sigma_1^{\mathrm{j}})^2},\cdots,\frac{(\sigma_{\mathrm{m}}^{\mathrm{j}})^2}{1+(\sigma_{\mathrm{m}}^{\mathrm{j}})^2}\right)\right)$
  5: Update $\sigma_i^j$
        (i) Compute $b_i$ the $i^{th}$ diagonal entry of $(\mathbf{U}_M^{j-1})'\mathbf{Y}'\mathbf{\Sigma}_x^2\mathbf{Y}(\mathbf{U}_M^{j-1})$.
        (ii) Compute $c_i$ the $i^{th}$ diagonal entry of $(\mathbf{U}_M^{j-1})'\mathbf{Y}'\mathbf{Y}(\mathbf{U}_M^{j-1})$
respectively.
        (iii) Solve: $\sigma_i^2=\left(\sqrt{\frac{b_i}{c_i v}}-1\right)^+$ and $\sum_{i=1}^m c_i\left(\sqrt{\frac{b_i}{c_i v}}-1\right)^+=\alpha^2$
6: Compute $\mathbf{M}=\mathbf{U}_M^N\mathbf{Diag}\left(\sigma_1^N,\cdots,\sigma_n^N\right)$
Output: $\mathbf{A}=(\mathbf{YM})'$
```

Table 5.1: Iterative algorithm for solving the sensing matrix design problem (5.4).

The sub-problem $\mathbf{P_0}$ is maximization over $\{\sigma_i\}_{i=1}^m$ for fixed $\mathbf{U}_M$, and $\mathbf{P_1}$ is maximization over $\mathbf{U}_M$ for fixed $\{\sigma_i\}_{i=1}^m$. The main novelty here is in the way we split the constraints. We next demonstrate how to solve these sub-problems.

5.6.1 Solving P0

With some linear algebra we can show **P0** is equivalent to

$$\begin{array}{ll} \underset{\gamma_i\in\mathbb{R}\forall i=1\ldots m}{\text{maximize}} & \sum_{i=1}^m \frac{b_i\gamma_i}{1+\gamma_i} \\ \text{subject to} & \sum_{i=1}^m c_i\gamma_i\leq\alpha^2,\ \gamma_i\geq 0, \quad i=1,\cdots m. \end{array} \tag{5.7}$$

where $\gamma_i=\sigma_i^2$, and b_i and c_i are the i^{th} diagonal entry of $\mathbf{U}_M'\mathbf{Y}'\mathbf{\Sigma}_x^2\mathbf{Y}\mathbf{U}_M$ and $\mathbf{U}_M'\mathbf{Y}'\mathbf{Y}\mathbf{U}_M$ respectively. With this, we can show that **P0** is a convex problem whose solution is given by

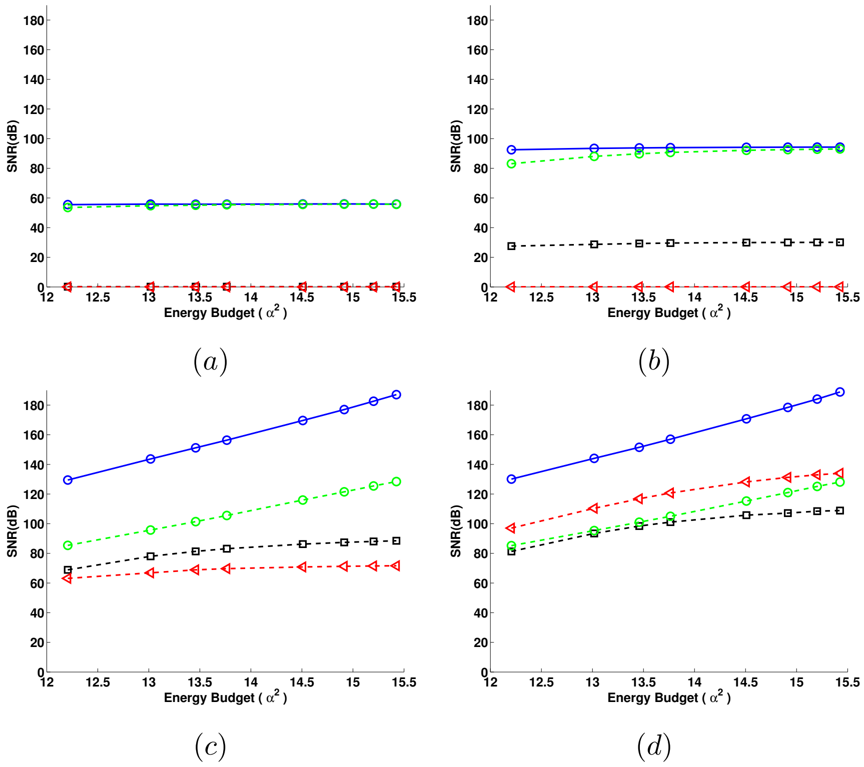
$$\gamma_i^* = \left(\sqrt{\frac{b_i}{c_i v^*}} - 1\right)^+ \quad \text{and} \quad \sum_{i=1}^{m} c_i \left(\sqrt{\frac{b_i}{c_i v^*}} - 1\right)^+ = \alpha^2, \tag{5.8}$$

where $(a)^+ = \max\{0, a\}$ and v^* is the Lagrangian multiplier associated with the constraint $\sum_{i=1}^{m} c_i \gamma_i \leq \alpha^2$ which can be solved by binary search algorithm similar to standard water-filling solution with a minor modification.

5.6.2 Solving P1

Owing to orthonormality constraints this is a challenging sub-problem. However, using the Lagrangian approach and with a carefully crafted argument along the lines of first order optimality conditions the problem **P1** can be effectively to converted to choosing m eigen vectors out of n eigen vectors of $\mathbf{Y}'\mathbf{\Sigma}_x^2\mathbf{Y}$ so that the i^{th} largest eigenvalue of $\mathbf{Y}'\mathbf{\Sigma}_x^2\mathbf{Y}$ is multiplied with i^{th} largest value in the set $\left\{\frac{\sigma_k^2}{1+\sigma_k^2}\right\}_{k=1}^{m}$. This gives us the optimal solution to **P1**. Details can be found in [6].

Based on the above solutions to sub-problems the final alternating minimization algorithm for solving (5.4) is shown in Table 5.1.



(a) (b) (c) (d)

Figure 5.1: Reconstruction SNR $= 20 \log \frac{\|\mathbf{x}\|_2}{\|\mathbf{x}-\hat{\mathbf{x}}\|_2}$ vs. sensing energy budget for several different compressive measurement strategies (see text for details). Panels (a)-(d) correspond to $m = 20, 40, 60, 80$ measurements, respectively. Higher SNR values correspond to better reconstructions. Our proposed approach (blue dotted line, circle markers) outperforms each of the other measurement strategies examined.

5.7 Experimental evaluation

We evaluate the performance of our proposed sensing matrix design procedure via experimentation on synthetic data. We consider signals of dimension $n = 100$, for which the number of signal and clutter models are $m_x = m_c = 10$, and where each model (in each class) is a covariance matrix of rank 6. The actual covariance matrices of the signal and clutter models are constructed randomly using a (different) random set of n orthonormal n-dimensional vectors, and randomly generated (positive) singular values.

For a subset of possible values of m we perform 1000 trials of the following experiment. First, we select one model randomly from the set $\{\mathbf{\Sigma}_{x,i}\}_{i=1}^{m_x}$ and generate $\mathbf{x}$ as a zero-mean Gaussian random vector having this covariance matrix, and we generate $\mathbf{c}$ similarly using one model selected randomly from $\{\mathbf{\Sigma}_{c,i}\}_{i=1}^{m_c}$. We then generate four different sets of observations $\mathbf{y}_i$ obtained using corresponding measurement matrices $\mathbf{A}_i$, for $i = 1, \ldots, 4$, as follows. First, For $\mathbf{C}$ an $m \times n$ matrix whose elements are iid $\mathcal{N}(0, 1/m)$ random variables, we let $\mathbf{A}_1 = \mathbf{C}(\alpha \mathbf{I}_n)$ denote observations obtained by traditional random projections. Next, we let $\mathbf{A}_2 = \mathbf{A}^*$, where $\mathbf{A}^*$ is the solution of (5.4) corresponds to the sensing matrix designed via our approach. We also compare with two more "heuristic" approaches – in the first of these, we form the sensing matrix $\mathbf{A}_3$ from a low rank approximation of the Wiener filter for estimating $\mathbf{x}$ from the mixture $\mathbf{x} + \mathbf{c}$, as discussed in [144]. Specifically, here we form $\mathbf{W}_{lr} = \mathbf{B}(\mathbf{\Sigma}_x + \mathbf{\Sigma}_c)^{-1/2}$ where $\mathbf{B}$ is the best rank m approximation of $\mathbf{\Sigma}_x(\mathbf{\Sigma}_x + \mathbf{\Sigma}_c)^{-1/2}$ (in the least-squares sense). Then, we let $\mathbf{W}_{lr} = \mathbf{U}_{\mathbf{W}_{lr}} \mathbf{\Sigma}_{\mathbf{W}_{lr}} \mathbf{V}'_{\mathbf{W}_{lr}}$ and obtain the sensing matrix $\mathbf{A}_3$ by retaining the first m rows of the matrix $\mathbf{\Sigma}_{\mathbf{W}_{lr}} \mathbf{V}'_{\mathbf{W}_{lr}}$, and appropriately rescaling to meet the sensing energy constraint. This represents a case where we employ a classic (linear) estimation strategy directly into the measurement process while keeping in mind that we are allowed only to take m measurements. We also investigate the estimation performance associated with the sensing matrix $\mathbf{A}_4 = \check{\mathbf{A}}^*$, where $\check{\mathbf{A}}^*$ represents the solution of (5.4) in a modified setting where clutter models are not viewed as clutter, but rather as additional *signal* models. In words, this describes the case where we design $\mathbf{A}$ in order to accurately estimate the mixture $\mathbf{x} + \mathbf{c}$, deferring the separation entirely to a subsequent step. The additive noise in each case is $\mathbf{w} \sim \mathcal{N}(0, \mathbf{I}_{m \times m})$.

We aim to reconstruct the signal $\mathbf{x}$ in each case using a group lasso approach [145] that explicitly leverages the correlation structure described by each model. To that end, we let $\mathbf{D}_x$ be the $n \times (6 \cdot m_x)$ signal dictionary whose $n \times 6$ blocks correspond to the top 6 eigenvectors of each of the signal models $\mathbf{\Sigma}_{x,1}, \ldots, \mathbf{\Sigma}_{x,m_x}$, and similarly for $\mathbf{D}_c$, and we denote by $\mathbf{D} = [\mathbf{D}_x \ \mathbf{D}_c]$ the combined dictionary, comprised of a total of $m_x + m_c$ models. Then, we obtain the estimates for each sensing matrix/observation vector pair as $\widehat{\mathbf{x}}_i = [\mathbf{D}_x \ \mathbf{0}] \ \left[\arg\min_\beta \|\mathbf{y}_i - \mathbf{A}_i \mathbf{D}\beta\|_2^2 + \lambda\Omega(\beta)\right]$, for $i = 1, \ldots, 4$ where the parameter vector β is $6 \cdot (m_x + m_c) \times 1$ and the regularizer $\Omega(\beta) = \sum_{j=1}^{m_x+m_c} \sqrt{\mathbf{v}'_j \mathbf{\Lambda}_j^{-1} \mathbf{v}_j}$, where $\mathbf{v}_j = \beta_{[6(j-1)+1:6j]}$ is a sub vector of the parameter vector corresponding to the

j-th overall model and $\mathbf{\Lambda}_j$ is the 6×6 diagonal matrix whose elements are the nonzero eigenvalues of the j-th model. Optimizations were performed using the Sparse Modeling Software (SpaMS) [2] . We compare the performance of each of the four approaches in terms of reconstruction SNR vs. sensing energy budget α^2. The results, depicted in Figure 5.1, show that our proposed approach (blue line, circle marker) outperforms each of the other approaches – traditional CS (black dotted line, square markers), the sensing approach based on the low rank Wiener filter (green dotted line, circle markers), and the approach where the clutter models are viewed as signal models and separation is left to the final estimation step (red dotted line, triangle markers) – across all sensing energy budgets, and for each subsampling case examined ($m = 20, 40, 60, 80$ measurements, respectively, in panels (a)-(d)).

5.8 Summary

It is interesting to see that both our proposed approach as well as the low rank Wiener filter approach are performing a kind of "annihilate-then-estimate" sensing strategy, while the approach corresponding to the sensing matrix $\mathbf{A}_4$ is more of an "estimate-then-annihilate" strategy. Our results here suggest that the former approach is more viable here – in other words, our empirical results here suggest that we should incorporate some "cancellation" into the sensing matrix itself, rather than relying on the final estimation step to perform the separation. Further, while our design approach was based on a MSE minimization criteria, we note a point of comparison between our approach and related design strategies that are based on maximizing mutual information between the vector $\mathbf{x}$ to be estimated and observations obtained for a specific $\mathbf{A}$. At first glance, these criteria are (seemingly) different, however, a fundamental connection between the minimum MSE matrix and the mutual information between the unknown $\mathbf{x}$ and the observations $\mathbf{y}$ (more specifically, its gradient with respect to various problem parameters, such as the matrix $\mathbf{A}$) has recently been established – see [146]. Indeed, the work [146] discusses a related task of linear per-coder design in an effectively "clutter-free" scenario, and proposes a gradient projection approach for obtaining the optimal precoder matrix.

[2] Available online at `http://spams-devel.gforge.inria.fr`

Chapter 6

A compressed sensing based decomposition of electrodermal activity signals

Electrodermal Activity, or EDA, is typically recorded as the conductance over a person's skin, near concentrations of sweat glands (*e.g.,* palm of the hand or finger tips [147]). EDA signals have been shown to include significant information pertaining to human neuron firing [148] and psychological arousal [149].[1] While previously a signal that was only practically measured in a controlled laboratory setting, recent wearable devices, such as the Affectiva Q sensor [150] and the Empatica E4 sensor [151], offer the ability to non-invasively measure EDA signals in real-world environments.

An EDA signal is generally characterized by a slowly changing Skin Conductance Level (SCL) combined with several short-lived Skin Conductance Responses (SCRs). The physiological explanation can be summarized as follows: the SCL is measuring the overall absorption of sweat in the user's skin, while each SCR is measuring a discrete event of sweat expulsion triggered by user excitement or psychological arousal in response to stimuli [152]. We refer to these discrete events as *SCR events.* The primary focus of prior EDA signal analysis has been to extract the informative SCR events

[1] The material in this chapter is ©2016 IEEE. Reprinted, with permission, from *IEEE Transactions on Biomedical Engineering*, "A Compressed Sensing Based Decomposition of Electrodermal Activity Signals", S. Jain, U. Oswal, K. Xu, B. Eriksson, and J. Haupt.

from the observed signals, due to applications ranging from content valence classification [153], to audience cohort analysis [154], to stress detection [155]. This can prove to be quite challenging due to the overlap of SCR signal components, a dominant SCL signal, signal artifacts due to motion, and the inclusion of measurement noise. As a result, there are a large number of proposed techniques to extract SCR events from observed EDA signals [152, 153, 156–160], which are discussed in the ensuing sections of this chapter.

Unfortunately, these prior techniques have a series of drawbacks. First, many of these techniques perform only simple heuristic-based approaches to extract the SCR events, which causes the techniques to be sensitive to noise and motion artifacts, *i.e.* sudden shifts in skin conductance due to changes in the position of the sensor. Second, these techniques lack error bounds on the recovered SCR events, so there is no guarantee for accuracy. Finally, most of the prior methods have ignored the contribution of motion artifacts. As EDA becomes more commonly observed via wearable devices, it is more important to mitigate such motion artifacts.

Here, we offer a new, more realistic EDA signal model that considers the observed EDA signal as the superposition of a *baseline* signal (signal component due to SCL changes and motion artifacts), informative SCR components, and measurement noise. Given this cluttered observed signal, we discuss how existing signal de-mixing work (*e.g.,* [161, 162]) indicates significant challenges in reliably extracting our desired sparse SCR event signal. We overcome these challenges by providing a new signal model for the baseline signal component which captures changes in measured skin conductance due to motion as well as changes in SCL. Further, we exploit this signal structure by a simple pre-processing step, which transforms this recovery problem into the more tractable problem of sparse deconvolution in the presence of bounded noise.

The problem of sparse deconvolution has been examined extensively in the compressed sensing literature (*e.g.,* [52, 53, 55, 163–165]). We show how our EDA problem setup requires additional changes to the standard compressed sensing problem. We use modified compressed sensing tools to estimate the SCR events using a concise optimization program and corresponding recovery error bounds. This results in “first-of-its-kind” EDA signal decomposition with known error rates.

We test this methodology on a series of both synthetic and real-world EDA signals.

Using synthesized data we are able to sweep varying noise and sparsity levels to reveal regimes where our technique accurately recovers the sparse responses. We then show on real-world EDA signals that user reactions to simple stimuli can be extracted with high accuracy compared with existing EDA decomposition algorithms.

6.1 Related work

The study of Electrodermal Activity signals, or EDA signals, dates back to the early 20$^{\text{th}}$ century (*e.g.,* [149]) with the observation of a connection between changes in user skin conductance and psychological state. In recent years, this connection has been validated by examining brain function via fMRI and skin conduction via EDA concurrently in [166], and by showing the specific regions of the brain that correspond with EDA changes and video recordings of sweat glands in [148]. The promise of EDA as a window into user psychology resulted in extensive work on evaluating the connection between EDA and user interactions [167], stress detection [155], content and audience segmentation [154], and reaction to video content [153]—to name only a few.

Applications using EDA signal analysis rely on the extraction of a user's fine-grained responses embedded in the EDA signal called Skin Conductance Responses, or SCRs. These SCRs measure the expulsion of sweat triggered by a user's spike-like stimulus responses, which we call SCR events. SCR events are not explicitly observed in the EDA signal; we observe only the SCRs, which can be modeled as the convolution of the SCR events with a distinguishing impulse response. Significant prior literature has focused both on how to model the SCR impulse response and extract the SCR events from the observed EDA signal. Examples include a parametric sigmoid-exponential model [156], a bi-exponential impulse response [157], nonnegative deconvolution [152], and a variational Bayesian decomposition methodology [158]. These prior techniques are limited by either computational complexity [158] or overly simple models that ignore or heuristically remove additional EDA signal components, such as the SCL, that disguise the SCR events [152, 157].

The authors of [152] treat the SCL as a constant estimated by averaging the skin conductance signal over the time windows when the estimated SCR (by deconvolution) is below a certain amplitude. The work of [153] presented a methodology to extract relevant SCR events while considering the SCL signal, but their matching pursuit-based technique used only a rough heuristic to remove this additional signal by deleting the two coarsest-scale components of a discrete-cosine transform applied to the skin conductance.

More recent work has incorporated SCL in a more principled manner into the EDA signal model. The sparse representation of SCR signal was exploited in [160]. In this work, the SCL signal was modeled as a slowly varying linear signal, and the SCR signal was modeled as a sparse linear combination of atoms of a dictionary containing time shifts of variety of function shapes. A greedy method exploiting the sparsity was also proposed for extracting the SCR events signal. Recently, the authors of [159] proposed an approach which exploited sparsity from a Bayesian perspective in which the SCL signal was modeled as a sum of cubic B-spline functions, an offset and a linear trend, whereas the SCR signal was modeled by a sparse signal in the dictionary obtained by shifts of bilinear transformations of a Bateman function. Following the maximum *a posteriori* (MAP) estimation principle, a convex formulation was obtained which can be solved efficiently. In contrast to these works [159, 160] we propose a model for the baseline signal that incorporates shifts in skin conductance due to changes in the positioning of the sensors due to motion, which is crucial when data is collected using wearables.

Given the sparse nature of the SCR events signal, in order to obtain bounds on our recovery, we leverage literature on compressed sensing [163]. Usually focused on sparse signal inference after transformation by random sensing matrices, here we are informed by recent work on sparse deconvolution in a compressed sensing regime [53], de-mixing of structured signals [161,162], and corrupted sensing for signals with known structure [168]. Our analysis differs from this prior work via the inclusion of a baseline signal model. This requires significant reformulation of the problem to develop new theory and recovery methodologies.

6.2 Model

The observed EDA skin conductance signal is typically characterized by two dominant components. The first is a slowly varying Skin Conductance Level (SCL), also referred to as the “tonic” component. The second component is the observation of multiple Skin Conductance Responses (SCRs) arising each from a corresponding SCR event. This component is sometimes referred to as the “phasic” component. These two signal types are detailed in Figure 6.1.

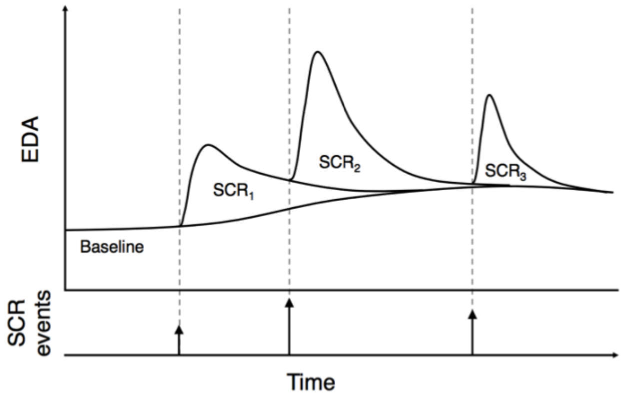


Figure 6.1: An example of EDA signal where the Skin Conductance Responses (SCRs) resulting from SCR events signal are shown.

The user’s physiology explains the existence of these two signal components. The SCRs are driven by occurrences of SCR events, a sparse selection of events where the user has responded with psychological arousal or excitement to stimulus. The SCR events signal is denoted by the impulse train at the bottom of Figure 6.1. Prior research in the psychophysiology community (*e.g.,* [152]) has recognized that these SCR events (*i.e.,* user excitement events) are correlated with sudomotor neuron bursts, resulting in a user’s eccrine glands to expel sweat. This sweat causes changes in skin conductance in the form of an SCR observation in the shape similar to that shown in Figure 6.1. This shape is the result of expelling, pooling, and evaporation of sweat on the surface of the user’s skin.

Additionally, this act results in some sweat being absorbed into the surface of the user’s skin, which affects the SCL. We consider the SCL to be a slowly varying signal. The SCL signal can be changed by temperature, humidity, and other environmental factors along with the physiology of the user (*e.g.,* thickness of the user’s skin).

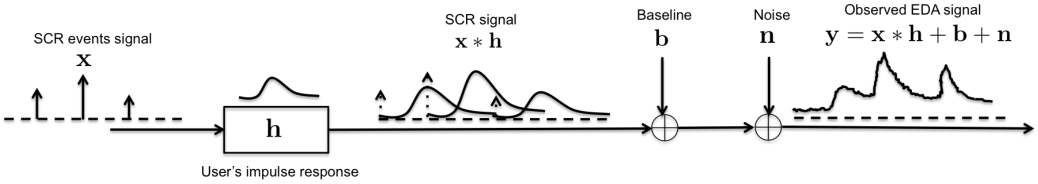


Figure 6.2: Observation model showing the various components in the observed EDA signal.

In addition to the SCL, there may also be sudden shifts in the skin conductance caused by changes in the positioning of the sensors or the amount of contact of the sensors with the skin, especially in the wearable sensor setting. Such changes are often reflected by jump discontinuities in the skin conductance. We account for such discontinuities, as well as the SCL, in what we call the *baseline* signal component.

6.2.1 Model definition

Let us consider an observed EDA signal, $\mathbf{y}$, discretized into T time steps. At each time step there is the possibility of an SCR event. We denote the SCR events signal corresponding to this content by a vector $\mathbf{x} \in \mathbb{R}^T$, where each component represents the intensity of the user's reaction to the T possible events. Whenever the user has an SCR event, prior research has shown (*e.g.,* [152, 156, 157]) that there are typical ways in which the EDA measurements record conductance changes. We denote this typical sweat response of an user by a vector $\mathbf{h} \in \mathbb{R}^t$. In the past [152, 157], the resulting SCR signal has been modeled as a linear time-invariant (LTI) system where the SCR events signal $\mathbf{x}$ is convolved with the sweat response signal $\mathbf{h}$ which we denote as $\mathbf{h}*\mathbf{x} \in \mathbb{R}^{t+T-1}$.

As mentioned earlier, the SCR signal $\mathbf{h} * \mathbf{x}$ is superimposed with a baseline signal consisting of SCL and motion artifacts. Denote the baseline signal as $\mathbf{b} \in \mathbb{R}^{t+T-1}$ and the errors arising due to observation noise and model mismatch as $\mathbf{n} \in \mathbb{R}^{t+T-1}$. These notations are summarized in Table 6.1. The observed EDA signal can now be represented as

$$\mathbf{y} = \mathbf{h} * \mathbf{x} + \mathbf{b} + \mathbf{n}. \tag{6.1}$$

The final observation model is shown in Figure 6.2. Given prior work on the shape of the SCR impulse response $\mathbf{h}$, we assume that the impulse response is known *a priori* (we discuss the specific choice of $\mathbf{h}$ in Section 6.4). We consider the SCR events signal $\mathbf{x}$, the baseline $\mathbf{b}$, and noise $\mathbf{n}$ to all be unknown.

Table 6.1: EDA Signal Notation Summary

Component	Notation	Description
Baseline	$\mathbf{b}$	*Baseline Signal* - Slowly varying skin conductance level with jump discontinuities due to motion
SCR Events	$\mathbf{x}$	*Skin Conductance Response Events* - Signal of sparse stimulus response events from the user
SCR	$\mathbf{h} * \mathbf{x}$	*Skin Conductance Response* - Measured sweat expulsion resulting from the SCR events
Noise	$\mathbf{n}$	Additive noise observed from measurement process and model mismatch

We propose a model for the observed EDA signal $\mathbf{y}$ that accounts for both the baseline $\mathbf{b}$ and observation noise $\mathbf{n}$ in a principled manner. This requires further specifications on the signals $\mathbf{x}$, $\mathbf{b}$, and noise $\mathbf{n}$ which we detail in the following.

6.2.2 SCR events signal model

Due to physiology, there are limitations to how often humans can generate SCR events. Motivated by this, we impose a sparsity assumption on the SCR events signal. Specifically, we assume that there are no more than $s < T$ events to which a user responds significantly. More formally, the SCR events signal is assumed to lie in the set

$$\mathcal{X}_\delta^s = \left\{ \mathbf{x} \,\middle|\, \mathbf{x} \in \mathbb{R}^T, \|\mathbf{x} - \mathbf{x}_s\|_1 \leq \delta \right\}, \tag{6.2}$$

where δ is a small constant and $\mathbf{x}_s \in \mathbb{R}^T$ with exactly s non-zero components obtained by retaining the s-largest magnitude components of $\mathbf{x}$.

The above set is the collection of vectors which can be approximated within some distance (in terms of the ℓ_1-norm) δ from an exactly s-sparse signal. Notice that when $\delta = 0$, the above set is the set of s-sparse vectors in $\mathbb{R}^{t+T-1}$. We note that in most prior literature, the model for the SCR events signal is strictly positive. Here we drop this constraint for a simpler analysis of recovery guarantees. Our experimental results in Section 6.4 show that even without positivity constraints comparable performance can be achieved.

6.2.3 Baseline model

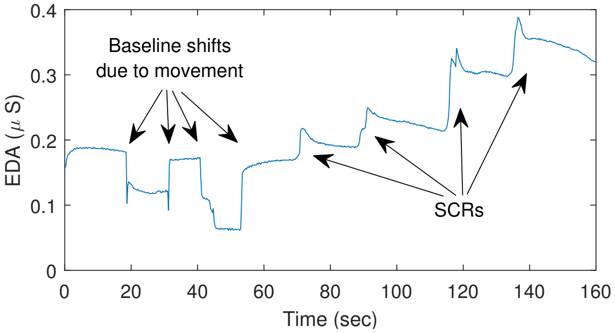


Figure 6.3: An example EDA signal collected using a commercially available wearable EDA sensor showing the impact of baseline shifts due to movement.

We propose a novel baseline model, inspired by the wearable setting where changes in the positions of sensors due to movement may lead to rapid changes in the EDA signal. These rapid changes, or baseline shifts, are illustrated in Figure 6.3 along with several SCRs. To the best of our knowledge, such baseline shifts have not been examined by previous work on recovering SCR events. We incorporate these baseline shifts along with the SCL component into a baseline signal $\mathbf{b}$. We assume $\mathbf{b}$ changes its magnitude significantly or has jump discontinuities at no more than $c < t + T - 1$ locations. More formally, the baseline signal is assumed to lie in the set

$$\mathcal{B}_\gamma^c = \left\{ \mathbf{b} \,\middle|\, \mathbf{b} \in \mathbb{R}^{t+T-1}, \|\mathbf{Db} - (\mathbf{Db})_c\|_1 \leq \gamma \right\}, \tag{6.3}$$

where $\mathbf{D} \in \mathbb{R}^{(t+T-2)\times(t+T-1)}$ denotes the pairwise difference matrix defined by

$$\mathbf{D} = \begin{bmatrix} 1 & -1 & 0 & \cdots & 0 \\ 0 & 1 & -1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 & -1 \end{bmatrix} \tag{6.4}$$

so that $\mathbf{Db} = [b_1 - b_2, b_2 - b_3, \ldots, b_{t+T-2} - b_{t+T-1}]$ and $(\mathbf{Db})_c \in \mathbb{R}^{t+T-2}$ with exactly c non-zero components obtained by retaining the c-largest magnitude components of $\mathbf{Db}$. Hence the baseline signal, after pairwise differencing, is assumed to be within some distance (in terms of the ℓ_1-norm) γ from a c-sparse signal.

6.2.4 Bounded noise model

Finally, we consider the additional noise induced by the wearable sensor recording the EDA signals as well as potential model mismatch. Rather than assuming a form for the distribution of this term, we will simply assume that the noise and model inaccuracies are bounded by a fixed value, *i.e.* $\|\mathbf{n}\|_2 \leq \epsilon/2$ where $\epsilon > 0$. Here the constant factor $1/2$ is included only to simplify further analysis.

6.2.5 Problem overview

The goal is to obtain the SCR events signal $\mathbf{x}$ from the EDA observation signal $\mathbf{y} = \mathbf{h} * \mathbf{x} + \mathbf{b} + \mathbf{n}$ given the prior information that $\mathbf{x} \in \mathcal{X}_\delta^s$ and $\mathbf{b} \in \mathcal{B}_\gamma^c$. We assume that the impulse response $\mathbf{h}$ is known, but the baseline $\mathbf{b}$, the SCR events signal $\mathbf{x}$, and the measurement noise $\mathbf{n}$ are all unknown.

6.3 EDA signal decomposition

The task of recovering the true SCR events $\mathbf{x}$ from the observed EDA signal $\mathbf{y}$ is particularly challenging due to the presence of the baseline $\mathbf{b}$. For example, consider the setting when there is an observed signal with no baseline and no noise, *i.e.,* $\mathbf{b} = \mathbf{0}$, $\mathbf{n} = \mathbf{0}$, and $\mathbf{y} = \mathbf{h} * \mathbf{x}$. The problem of recovering $\mathbf{x}$ from $\mathbf{y}$ simply reduces to solving an over-determined linear system of equations given knowledge of $\mathbf{h}$. As a result, this problem can be solved with standard deconvolution techniques given very mild assumptions on $\mathbf{h}$ and without any assumptions needed on true $\mathbf{x}$.

In another case, consider there is no baseline but noise is present, *i.e.,* $\mathbf{b} = \mathbf{0}$, $\mathbf{n} \neq \mathbf{0}$, and $\mathbf{y} = \mathbf{h} * \mathbf{x} + \mathbf{n}$. This is a standard problem of deconvolution in noise, which in general is a difficult problem to solve. But, when we consider the added structure of the sparsity of SCR events signal $\mathbf{x}$, one could exploit this to estimate $\mathbf{x}$ with provable guarantees. This setting has been explored in prior work in the field of compressed sensing, *e.g.,* [53].

6.3.1 Dealing with the baseline signal

The main challenge here is the case where the baseline signal is present and non-zero. One obvious approach could be to consider the baseline as noise and follow previously proposed deconvolution for noisy settings *e.g.,* [53]. However, this would likely fail because the baseline $\mathbf{b}$ could have very large magnitude. Our proposed alternative is to exploit the structure of the baseline signal to facilitate the recovery of $\mathbf{x}$. We linearly transform the baseline signal and jointly recover the transformed baseline and $\mathbf{x}$. This is often known as a *de-mixing* problem, and there has been recent work on using convex techniques for de-mixing structured signals [161, 162]. These papers have theoretical guarantees in terms of statistical dimension. Unfortunately, these guarantees assume a specific random signal generation model which does not hold true for our problem setting.

Recent work has proposed a *corrupted sensing* approach [168] which extends compressed sensing to a setting where observations are corrupted with structured signals. Our problem is different from this setup on two counts: (1) Our sparse signal is convolved with a known SCR impulse response and (2) the baseline signal in our setting has structure that has not yet been considered in the corrupted sensing literature. Hence, we leave this as an interesting future direction.

6.3.2 EDA signal preprocessing

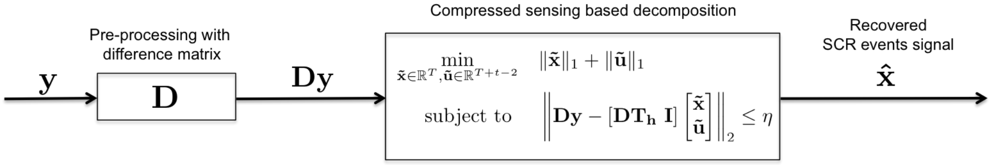

Compressed sensing based decomposition
Pre-processing with difference matrix
$\mathbf{y}$
$\mathbf{D}$
$\mathbf{Dy}$
$\min_{\tilde{\mathbf{x}}\in\mathbb{R}^T, \tilde{\mathbf{u}}\in\mathbb{R}^{T+t-2}} \quad \|\tilde{\mathbf{x}}\|_1 + \|\tilde{\mathbf{u}}\|_1$
subject to $\left\| \mathbf{Dy} - [\mathbf{DT_h} \ \mathbf{I}] \begin{bmatrix} \tilde{\mathbf{x}} \\ \tilde{\mathbf{u}} \end{bmatrix} \right\|_2 \leq \eta$
Recovered SCR events signal
$\hat{\mathbf{x}}$

Figure 6.4: Block diagram showing the SCR events signal recovery using compressed sensing based decomposition.

We propose an approach that exploits the structure of the EDA signals to mitigate the effects of the baseline signal. Namely, we can consider that the baseline signals have almost the same consecutive components for most of the signal elements. As a result, they can be converted to approximately sparse signals by multiplying with the pairwise difference matrix $\mathbf{D}$ defined in (6.4).

Of course, we only have access to the observed signal, $\mathbf{y}$. Therefore, we follow a very simple approach in which we linearly transform the observation $\mathbf{y}$ using the difference matrix $\mathbf{D}$ as follows:

$$\mathbf{Dy} = \mathbf{DT_h x} + \mathbf{Db} + \mathbf{Dn}, \tag{6.5}$$

where $\mathbf{T_h}$ denotes a $(t + T - 1) \times T$ Toeplitz matrix constructed from a vector $\mathbf{h} \in \mathbb{R}^t$ and is defined as follows:

$$\mathbf{T_h} = \underbrace{\begin{bmatrix} h_1 & 0 & \cdots & 0 \\ h_2 & h_1 & \vdots & \vdots \\ \vdots & \vdots & \ddots & 0 \\ h_t & h_{t-1} & \vdots & h_1 \\ 0 & h_t & & \\ \vdots & \vdots & \ddots & \\ 0 & \cdots & \cdots & h_t \end{bmatrix}}_{T \text{ columns}} \tag{6.6}$$

such that the convolution between vectors $\mathbf{h} \in \mathbb{R}^t$ and $\mathbf{x} \in \mathbb{R}^T$, denoted by $\mathbf{h} * \mathbf{x}$, is a vector in $\mathbb{R}^{t+T-1}$ and can be written in terms of matrix-vector multiplications as $\mathbf{h} * \mathbf{x} = \mathbf{T_h x}$.

With this transformation, the modified baseline signal $\mathbf{Db}$ is approximately sparse because of the structure of $\mathbf{b} \in \mathcal{B}_\gamma^c$. Due to this sparsity, the transformed baseline signal has similar structure to the true SCR events signal $\mathbf{x}$. We leverage this fact to jointly estimate $\mathbf{x}$ and $\mathbf{Db}$. Rearranging this term, the observation model becomes

$$\mathbf{Dy} = \begin{bmatrix} \mathbf{DT_h} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{x} \\ \mathbf{Db} \end{bmatrix} + \mathbf{Dn},$$

where $\mathbf{I}$ denotes the identity matrix. We have transformed this problem into estimating a vector that is approximately sparse with $s+c$ significant components in $\mathbb{R}^{t+T-2}$, where s is the number of significant non-zero elements in $\mathbf{x}$, and c is the number of significant non-zeros in $\mathbf{Db}$.

Using recent advances in compressed sensing [169], we propose to solve the following problem to estimate $\mathbf{x}$ and $\mathbf{Db}$:

$$\begin{aligned} \min_{\tilde{\mathbf{x}}\in\mathbb{R}^T,\tilde{\mathbf{u}}} \in \mathbb{R}^{T+t-2} \quad & \|\tilde{\mathbf{x}}\|_1 + \|\tilde{\mathbf{u}}\|_1 \\ \text{subject to} \quad & \left\| \mathbf{Dy} - [\mathbf{DT_h}\ \mathbf{I}] \begin{bmatrix} \tilde{\mathbf{x}} \\ \tilde{\mathbf{u}} \end{bmatrix} \right\|_2 \leq \eta, \end{aligned} \tag{6.7}$$

where $\eta > 0$ is a parameter that can be chosen based on the energy of noise $\mathbf{n}$ as detailed in the next subsection. The above problem is known to be a convex problem which can be solved by using well-known convex optimization software (*e.g.,* CVX [170]). The final recovery procedure based on above discussion is summarized in Figure 6.4. We note that our problem has Toeplitz structure which can be exploited for developing computationally efficient algorithm using the ideas from matrix-free convex optimization modeling [171, 172]. We leave this as an interesting future direction of work.

6.3.3 Error guarantees

The fundamental question that arises here is how well the estimates obtained by solving above problem work. Specifically, how close is the optimal solution $\hat{\mathbf{x}}$ of (6.7) to the true SCR events signal $\mathbf{x}$? We have the following theorem to specifically detail the error in our recovered SCR events signal.

Theorem 6.3.1. *Let* $\mathbf{y} = \mathbf{h} * \mathbf{x} + \mathbf{b} + \mathbf{n}$*, where* $\mathbf{x} \in \mathcal{X}_\delta^s, \mathbf{b} \in \mathcal{B}_\gamma^c$*. Denote* $\mathbf{C} = [\mathbf{DT_h}\ \mathbf{I}]$ *and define the coherence parameters* μ_h, μ_m, μ_c *as*

$$\mu_h = \max_{i \neq j} \frac{|\mathbf{t}_i^T \mathbf{t}_j|}{\|\mathbf{t}_i\|_2 \|\mathbf{t}_j\|_2}, \ \mu_c = \max_{i \neq j} \frac{|\mathbf{c}_i^T \mathbf{c}_j|}{\|\mathbf{c}_i\|_2 \|\mathbf{c}_j\|_2}$$

$$\mu_m = \max_{i,j} \frac{|\mathbf{t}_i^T \mathbf{e}_j|}{\|\mathbf{t}_i\|_2 \|\mathbf{e}_j\|_2}$$

where $\mathbf{t}_i, \mathbf{e}_i$*, and* $\mathbf{c}_i$ *are the* i^{th} *columns of matrices* $\mathbf{DT_h}, \mathbf{I}$*, and* $\mathbf{C}$*, respectively. If* $\|\mathbf{n}\| \leq \epsilon/2$ *and*

$$s + c < max \left\{ \frac{2(1+\mu_h)}{\mu_h + 2\mu_c + \sqrt{\mu_h^2 + \mu_m^2}}, \frac{1+\mu_c}{2\mu_c} \right\},$$

then the solution $\hat{\mathbf{x}}, \hat{\mathbf{u}}$ *of* (6.7) *using* $\epsilon \leq \eta$ *satisfies*

$$\|\mathbf{x} - \hat{\mathbf{x}}\|_2 \leq C_1(\epsilon + \eta) + C_2(\delta + \gamma)$$

where $C_1, C_2 > 0$ *depend on* μ_c*,* μ_h*,* μ_m*,* s*, and* c*.*

Proof. See Appendix. □

The above theorem states that, when the combined sparsity of the true SCR events signal and the baseline signal after the difference filter is small enough, the estimate of the SCR events signal $\hat{\mathbf{x}}$ is accurate. More specifically, the ℓ_2 norm of the error vector (*i.e.,* the difference between the true and the estimated SCR events signal) is upper bounded by a quantity which is proportional to the constants ϵ, δ and γ, which are part of our signal model, and the optimization parameter η, provided that it is chosen to be greater than or equal to ϵ. As long as these constants are small, our approach yields an accurate solution. In our setting, it is reasonable to assume that these constants are indeed small for the following reasons. The SCR events signal $\mathbf{x}$ is sparse due to physiological reasons, as previously discussed. The baseline signal should not have too many jump discontinuities provided that the user is not constantly moving the sensor, which causes $\mathbf{Db}$ to also be sparse. Finally, ϵ depends on the noise power and model mismatch and is small provided that the noise power is much lower than the signal power and that our model assumptions are close to reality.

The terms C_1 and C_2 are known to decrease with decreasing s, c [169]. This implies that the error in the recovery decreases as the signals become more sparse. The range of values of $s + c$ for which the error bounds holds depends on the coherence parameters. These parameters critically depend on the shape and length of $\mathbf{h}$ which we assume are known. It is known that with decreasing coherence parameters μ_c, μ_h, and μ_m, the recovery of a sparse signal improves [169]. All the coherence parameters can be viewed as the maximum entries of the sub-blocks of the matrix

$$\begin{aligned}\mathbf{G} &= \begin{bmatrix} (\mathbf{D}\mathbf{T_h}\boldsymbol{\Lambda})^T \\ \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{D}\mathbf{T_h}\boldsymbol{\Lambda} & \mathbf{I} \end{bmatrix} - \mathbf{I} \\ &= \begin{bmatrix} (\mathbf{D}\mathbf{T_h}\boldsymbol{\Lambda})^T \mathbf{D}\mathbf{T_h}\boldsymbol{\Lambda} - \mathbf{I} & (\mathbf{D}\mathbf{T_h}\boldsymbol{\Lambda})^T \\ \mathbf{D}\mathbf{T_h}\boldsymbol{\Lambda} & \mathbf{0} \end{bmatrix},\end{aligned}$$

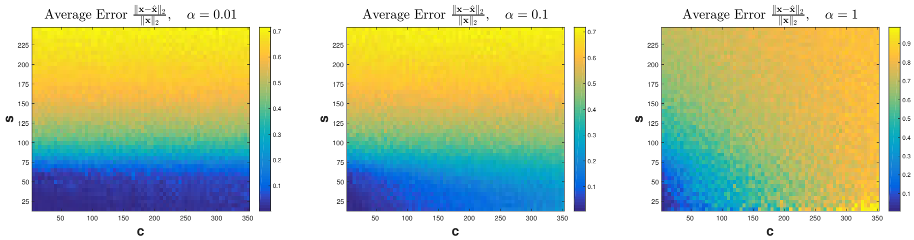
where $\boldsymbol{\Lambda}$ is a diagonal matrix such that the columns of the matrix $\mathbf{D}\mathbf{T_h}\boldsymbol{\Lambda}$ have unit ℓ_2 norm. The coherence parameters can be written in terms of sub-blocks of matrix $\mathbf{G}$ as follows

$$\begin{aligned}\mu_h &= \|(\mathbf{D}\mathbf{T_h}\boldsymbol{\Lambda})^T \mathbf{D}\mathbf{T_h}\boldsymbol{\Lambda} - \mathbf{I}\|_{\max} \\ \mu_m &= \|\mathbf{D}\mathbf{T_h}\boldsymbol{\Lambda}\|_{\max} \\ \mu_c &= \max\{\mu_h, \mu_m\},\end{aligned}$$

where for a matrix $\mathbf{X}$, the maximum absolute entry of the matrix is denoted by $\|\mathbf{X}\|_{\max}$.

6.4 Experiments

Using a combination of both synthetic and real-world EDA data, in this section we demonstrate the feasibility and accuracy of our proposed compressed sensing approach to EDA decomposition. Our synthetic data experiments sweep a wide selection of sparsity values and baseline signal energy levels to demonstrate SCR event recovery accuracy. Using real-world EDA data, we then show how our technique allows for more accurate inference of EDA events signal as compared to prior techniques.



(a) (b) (c)

Figure 6.5: Estimation error diagrams with synthetic data for various values of number of SCR events $s \in \{5, 10, \ldots, 230\}$ and baseline jumps $c \in \{5, 10, \ldots, 350\}$. Panels (a), (b), and (c) correspond to scaling the magnitude of the baseline component using $\alpha = 0.01, 0.1$ and 1, respectively.

6.4.1 Synthetic data experiment

The first experiment is dedicated to demonstrating the recovery accuracy of our procedure on synthetic data. We obtained the impulse response vector $\mathbf{h}$ by sampling the function $f(u)$ shown in Figure 6.6 at the rate of 4 samples per second in the interval $u \in [0, 40]$. This choice of impulse response was informed by prior psychophysiology literature [157]. The $\mathbf{h}$ obtained in such manner lies in $\mathbb{R}^{160}$. We fixed $T = 240, \delta = 0.01$ and $\gamma = 0.01$.

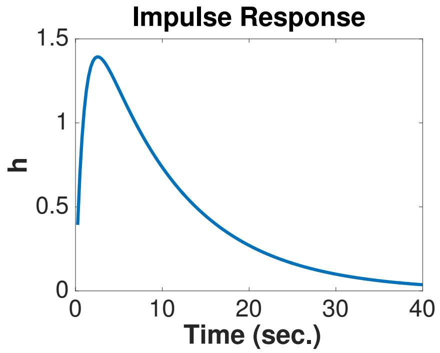
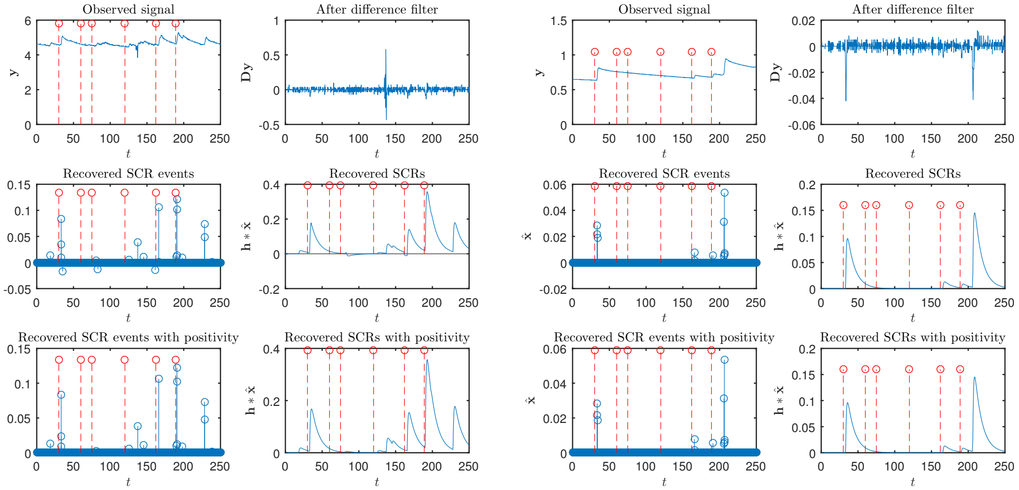


Figure 6.6: The impulse response $\mathbf{h}$ was obtained by sampling the function $f(u) = 2\left(e^{-\frac{u}{\tau_1}} - e^{-\frac{u}{\tau_2}}\right)$ for $u \geq 0$ and $f(u) = 0$ otherwise. Here $\tau_1 = 10, \tau_2 = 1$ and the is function sampled at the rate of 4 samples per second in the interval $u \in [0, 40]$.



(a) (b)

Figure 6.7: Decomposition of real-world EDA data for two users in (a) and (b) respectively. Stimuli are presented to the users at moments denoted by red dotted vertical lines. We show results for our compressed sensing approach with and without positivity constraints for data downsampled to 4 Hz.

For a given number of SCR events s and number of baseline jumps c, we randomly generate $\mathbf{x} \in \mathcal{X}_\delta^s$ and $\mathbf{b} \in \mathcal{B}_\gamma^c$. A random $\mathbf{x} \in \mathcal{X}_\delta^s$ is generated by first choosing the s significant components uniformly at random and filling these components with a random vector in $\mathbb{R}^s$ with i.i.d. exponentially distributed entries with mean 2. This is followed by adding to it a rescaled standard Gaussian random vector in $\mathbb{R}^T$ with ℓ_1 norm δ. Similarly, a random $\mathbf{Db}$ was generated by first choosing the c significant components uniformly at random and filling each of these components with a standard Gaussian variable followed by adding a rescaled standard Gaussian random vector in $\mathbb{R}^{t+T-2}$ with ℓ_1 norm γ. Using these steps we generate the observations as follows: x

$$\mathbf{Dy} = \mathbf{DT_h x} + \alpha \mathbf{Db} + \mathbf{n}, \tag{6.8}$$

where $\mathbf{n}$ is also a rescaled Gaussian random vector with ℓ_2 norm equal to $\epsilon = 0.01$. We generate multiple experiments using different values of α, a scaling factor applied to $\mathbf{Db}$ relative to $\mathbf{DT_h x}$. These observations are then used to obtain the estimate $\hat{\mathbf{x}}$ by solving the problem in (6.7) with $\eta = 1.05\epsilon$.

Figure 6.5 shows the average relative estimation error $\frac{\|\mathbf{x}-\hat{\mathbf{x}}\|_2}{\|\mathbf{x}\|_2}$, where the average is obtained by 30 random observations for various values of s and c. For baseline components with low energy in Figure 6.5a, we find that the ability to recover is almost entirely dependent on the number of SCR events embedded in the generated EDA signal. Regardless of the number of baseline jumps, we find that for fewer than 75 SCR events in an EDA signal, we can accurately recover the SCR signal. On the other hand, as the energy in the baseline increases, as shown in Figures 6.5b and 6.5c, we find that a large number of jumps in the baseline signal can degrade our ability to accurately recover the SCR events.

6.4.2 Experiments with real-world EDA data

Our second experiment examines the performance of our methodology on real-world EDA signals. We used EDA signals from a simple video stimulus experiment, originally published in [153]. The video consists of six short stimulus clips (each lasting less than 10 seconds) with differing levels of complexity. Specifically, this video contains a baby crying sound, a gun shot sound, a dog barking sound, the image of a gun, and two short videos of a subject injuring themselves. This stimulus is interspersed with silence where no audio or video is presented to the user. The EDA data consists of EDA traces from nine subjects (6 male, 3 female, with ages ranging between 20 and 50 years old) who watched the same video content in a darkened environment. The EDA was recorded using the Affectiva Q Sensor (available at `http://www.affectiva.com/`) with sampling at 32 Hz.

Unlike with the synthetic data experiment, we cannot assess relative estimation error $\frac{\|\mathbf{x}-\hat{\mathbf{x}}\|_2}{\|\mathbf{x}\|_2}$ because we do know the magnitudes of the ground-truth SCR events $\mathbf{x}$. We do, however, know the times at which the stimulus clips and periods of silence were presented to the users. Very few SCR events should occur during the periods of silence, while many SCR events should occur during the stimulus clips, thus we can use these times to assess how well our EDA decomposition technique is able to detect SCR events. Specifically, we used 10 second windows around each stimulus and silence clip, and then aggregated the estimated SCR event coefficients between the start of the clip and the end of the clip. These aggregated values are then compared to a threshold to produce a binary decision as to whether SCR events are present in the time window. The impulse response vector $\mathbf{h}$ was obtained by sampling the function $f(u) = 2(e^{-\frac{u}{\tau_1}} - e^{-\frac{u}{\tau_2}})$ for $u \geq 0$ and $f(u) = 0$ otherwise. We chose $\tau_1 = 10, \tau_2 = 1$. For our proposed technique and defined $\mathbf{h}$, we obtained estimates of SCR events signal for each user by solving (6.7) with $\eta = 0.14$.

To evaluate our performance we use four alternative methodologies: (1) aggregated raw EDA signal for each user in the stimulus and silence time windows, (2) the non-negative deconvolution analysis technique of Benedek and Kaernbach [152] using the Ledalab software package (available at `http://www.ledalab.de/`), (3) the convex optimization approach cvxEDA proposed in [159], and (4) a modification of our approach with positivity constraint for the SCR events signal[2] . The raw EDA analysis will communicate if the mean EDA signal is informative with respect to our stimulus, while the deconvolution approach demonstrates EDA decomposition that ignores the prominent baseline signal. The cvxEDA approach will compare our proposed model with a recent EDA decomposition technique using convex optimization. The approach with positivity constraints will test whether including positivity constraints in our problem setup improves recovery accuracy.

We perform experiments on the original 32 Hz data as well as 4 Hz and 8 Hz downsampled versions, which are more in-line with the sampling rates of commercially available wearable sensors such as the Empatica E4 [151] and Microsoft Band 2(4 and 5 Hz, respectively). For cvxEDA, the same values $\tau_1 = 10$ and $\tau_2 = 1$ as for our approach were used [3] , whereas for Ledalab, τ_1 and τ_2 were automatically optimized by the software package.

Discussion of results: The result of signal decomposition on the 4 Hz downsampled signal is shown in Figure 6.7. In this figure we highlight the recovered signals with our approach and a modified version with positivity constraints on the SCR events signal. Figures 6.7a and 6.7b correspond to two different users that were chosen at random from our data set. Stimuli are presented to the users at moments denoted by red dotted vertical lines. We see that the recovered SCR events signal is similar for both techniques except for the events with small negative amplitudes when no positivity constraints are

[2] Specifically, we solve problem (6.7) with positivity constraint $\mathbf{x} \geq 0$.

[3] cvxEDA also requires specification of the sampling interval δ, which was set to 1/sampling frequency, and other parameters δ_0, α, and γ, which were set to the default values in the software package.

enforced. The reconstructed SCR signal $\mathbf{h} * \hat{\mathbf{x}}$ using both approaches are also shown. Overall, we find that our proposed approach performs similarly to its variation with positivity constraints.

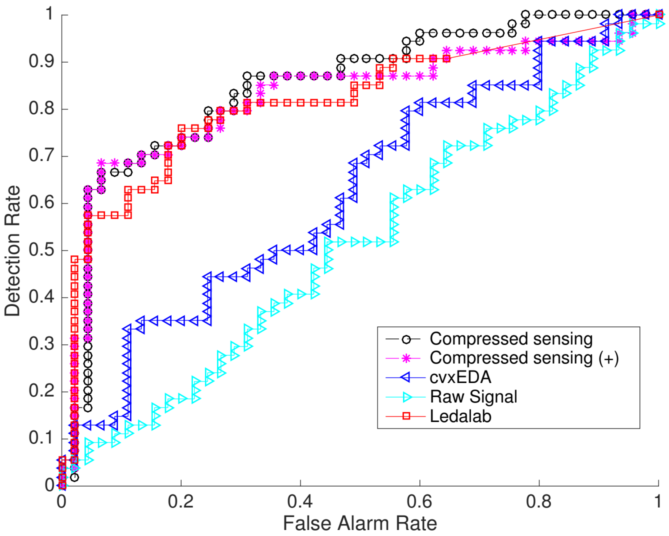


Figure 6.8: ROC curves for real data SCR event detection experiment at the sampling rate of 4 Hz. Our compressed sensing-based approaches is compared with a variation of our approach with positivity constraints, the non-negative deconvolution approach in Ledalab, the cvxEDA convex optimization based approach, and the raw EDA signal.

Further, aggregating the accuracy across all nine users, we present the Receiver Operating Characteristic (ROC) curve in Figure 6.8, which shows the detection rate for any given false alarm rate at the sampling rate of 4 Hz. We summarize the ROC curve using the Area Under the Curve (AUC). We find that our compressed sensing based decomposition (AUC = **0.848**) and its variation with positivity constraints (AUC = **0.825**) perform better than both the non-negative deconvolution method in Ledalab (AUC = **0.817**) and the convex optimization based cvxEDA approach (AUC = **0.622**). Another insight from these results is that using the raw EDA traces results in accuracy roughly no better than random guessing (i.e., detection rate equal to the false alarm rate), showing the need for processing of the observed EDA signals.

Table 6.2: AUC values for SCR event detection at multiple sampling rates for various approaches on real data experiment.

Sampling Rate	Compressed Sensing	Compressed Sensing (+)	cvxEDA	Raw Signal	Ledalab
4 Hz	**0.848**	0.825	0.622	0.539	0.817
8 Hz	**0.857**	0.821	0.771	0.493	0.824
32 Hz	0.868	**0.895**	0.819	0.514	0.837

The results at various sampling rates are shown in Table 6.2. We see that our scheme gives better performance than all other schemes at sampling rates 4 Hz and 8 Hz. This is an important regime when considering EDA observations from power and storage-constrained wearables. Our observations also suggest that, at these sampling rates, adding positivity constraints to our approach does not necessarily improve accuracy. In fact, at 4 Hz and 8 Hz, adding positivity constraints actually lowered the AUC. The only improvements for the positivity constrained techniques was at a sampling rate of 32 Hz.

6.5 Summary

In this work we proposed a novel compressed sensing based framework for processing of EDA signals. The proposed framework explicitly models the baseline signal and allows for recovery of the users responses via simple pre-processing followed by compressed sensing based decomposition. We also provided theoretical error bounds on the accuracy of the proposed recovery procedure. Our approach accurately recovers SCR events in experiments on simulated data. Furthermore, our recovery procedure also outperforms existing recovery procedures for an SCR event detection task on real-world EDA data obtained from a video stimulus experiment.

6.6 Appendix

Proof of Theorem 6.3.1. The proof is a straightforward extension of the following theorem from [169]:

Theorem 6.6.1 ([169], Thm. 4)**.** *Let* $\mathbf{t} = \mathbf{C}\mathbf{w} + \mathbf{z}$*, with* $\mathbf{C} = [\mathbf{A}\ \mathbf{B}]$*,* $\mathbf{w}^T = [\mathbf{x}^T\ \mathbf{u}^T]$*, and* $\|\mathbf{z}\|_2 \leq \epsilon$*. Define the coherence parameters* μ_a, μ_b, μ_m*, and* μ_c *for the dictionary* $\mathbf{C}$ *as*

$$\mu_a = \max_{i \neq j} \frac{|\mathbf{a}_i^T \mathbf{a}_j|}{\|\mathbf{a}_i\|_2 \|\mathbf{a}_j\|_2},\ \mu_b = \max_{i \neq j} \frac{|\mathbf{b}_i^T \mathbf{b}_j|}{\|\mathbf{b}_i\|_2 \|\mathbf{b}_j\|_2}$$
$$\mu_m = \max_{i,j} \frac{|\mathbf{a}_i^T \mathbf{b}_j|}{\|\mathbf{a}_i\|_2 \|\mathbf{b}_j\|_2},\ \mu_c = \max_{i \neq j} \frac{|\mathbf{c}_i^T \mathbf{c}_j|}{\|\mathbf{c}_i\|_2 \|\mathbf{c}_j\|_2}$$

Assume $\mu_b \leq \mu_a$ *without loss of generality. If*

$$s + c < max\left\{\frac{2(1+\mu_a)}{\mu_a + 2\mu_c + \sqrt{\mu_a^2 + \mu_m^2}}, \frac{1+\mu_c}{2\mu_c}\right\} \tag{6.9}$$

then the solution of $\hat{\mathbf{w}}$

$$\begin{aligned} &\min_{\tilde{\mathbf{w}}} && \|\tilde{\mathbf{w}}\|_1 \\ &\textit{subject to} && \|\mathbf{t} - \mathbf{C}\tilde{\mathbf{w}}\|_2 \leq \eta, \end{aligned} \tag{6.10}$$

using $\epsilon \leq \eta$ *satisfies*

$$\|\mathbf{w} - \hat{\mathbf{w}}\|_2 \leq C_1(\epsilon + \eta) + C_2\|\mathbf{w} - \mathbf{w}_{n+s}\|_1$$

where $C_1, C_2 > 0$ *depend on* μ_a, μ_b, μ_m*,* μ_c*, s, and c.*

We use the above Theorem 6.6.1 with $\mathbf{t} = \mathbf{D}\mathbf{y}, \mathbf{A} = \mathbf{D}\mathbf{T}_\mathbf{h}, \mathbf{B} = \mathbf{I}, \mathbf{z} = \mathbf{D}\mathbf{n}$, and $\mathbf{w}^T = [\mathbf{x}^T\ \mathbf{u}]$ with $\mathbf{u} = \mathbf{D}\mathbf{b}$. First we show that the ℓ_2 norm of the noise $\mathbf{z}$ satisfies the assumption in Theorem 6.6.1. This can be easily seen as follows

$$\begin{aligned} \|\mathbf{z}\|_2 &= \|\mathbf{D}\mathbf{n}\|_2 \\ &\leq \|\mathbf{D}\|_2 \|\mathbf{n}\|_2 \\ &\leq 2(\epsilon/2) = \epsilon, \end{aligned}$$

where the last inequality is due the fact that $\|\mathbf{D}\|_2 \leq 2$ and $\|\mathbf{n}\|_2 \leq \epsilon/2$ by our model assumption. Also, as $\mathbf{B} = \mathbf{I}$ is an orthonormal matrix, it is easy to see that $\mu_b = 0$.

Since μ_a is strictly positive under our model assumption, the condition $\mu_b \leq \mu_a$ is also satisfied. Further, since we can write $\|\tilde{\mathbf{w}}\|_1 = \|\tilde{\mathbf{x}}\|_1 + \|\tilde{\mathbf{u}}\|_1$, the optimization problem (6.10) in Theorem 6.6.1 takes the following form:

$$\begin{aligned} \min_{\tilde{\mathbf{x}} \in \mathbb{R}^T, \tilde{\mathbf{u}} \in \mathbb{R}^{T+t-2}} \quad & \|\tilde{\mathbf{x}}\|_1 + \|\tilde{\mathbf{u}}\|_1 \\ \text{subject to} \quad & \left\| \mathbf{Dy} - [\mathbf{DT_h} \ \mathbf{I}] \begin{bmatrix} \tilde{\mathbf{x}} \\ \tilde{\mathbf{u}} \end{bmatrix} \right\|_2 \leq \eta \end{aligned}$$

The above problem is exactly same as the problem in (6.7), for which error bounds are outlined in Theorem 6.3.1. This essentially establishes that Theorem 6.6.1 can be used to obtain the recovery guarantees of problem (6.7). Provided that the combined sparsity $s + c$ satisfies condition (6.9) and we choose η such that it satisfies $\epsilon \leq \eta$, we have the following bound from Theorem 6.6.1:

$$\begin{aligned} \|\mathbf{w} - \hat{\mathbf{w}}\|_2 &= \sqrt{\|\mathbf{x} - \hat{\mathbf{x}}\|_2^2 + \|\mathbf{Db} - \hat{\mathbf{u}}\|_2^2} \\ &\leq C_1(\epsilon + \eta) + C_2\|\mathbf{w} - \mathbf{w}_{n+s}\|_1 \\ &\leq C_1(\epsilon + \eta) + C_2\left\{\|\mathbf{x} - \mathbf{x}_s\|_1 + \|\mathbf{b} - \mathbf{b}_c\|_1\right\} \end{aligned}$$

Further, combining the above inequality with the fact that

$$\|\mathbf{w} - \hat{\mathbf{w}}\|_2 = \sqrt{\|\mathbf{x} - \hat{\mathbf{x}}\|_2^2 + \|\mathbf{Db} - \hat{\mathbf{u}}\|_2^2} \geq \|\mathbf{x} - \hat{\mathbf{x}}\|_2,$$

we have arrive at

$$\|\mathbf{x} - \hat{\mathbf{x}}\|_2 \leq C_1(\epsilon + \eta) + C_2\left\{\|\mathbf{x} - \mathbf{x}_s\|_1 + \|\mathbf{b} - \mathbf{b}_c\|_1\right\},$$

which, by our model assumption, can be reduced to

$$\|\mathbf{x} - \hat{\mathbf{x}}\|_2 \leq C_1(\epsilon + \eta) + C_2(\delta + \gamma).$$

The coherence parameters $\mu_h, \mu_m,$ and μ_c in Theorem 6.3.1 are equivalent to coherence parameters μ_a, μ_m, and μ_c respectively in Theorem 6.6.1. □

Chapter 7

Future Directions

7.1 Computationally-efficient approximations to arbitrary linear dimensionality reduction operators

So far our analysis has been focused on circulant structure. However, there are other structures like that are also known to have computational advantages. For example, in wireless communication the sparsity structure was exploited to control the backhaul traffic incurred due to the multicellular cooperation [8, 10]. An interesting direction of further study could be towards developing a similar fundamental understanding of computationally efficient approximations of LDR operators using other structured matrices.

A natural extension of our work on circulant structure based approximations of LDR operators is to extend the study to non-linear dimensionality reduction operators of the form defined as follows

$$f(\mathbf{x}) = \begin{bmatrix} g_1(\mathbf{a}_1^T\mathbf{x}) \\ \vdots \\ g_m(\mathbf{a}_m^T\mathbf{x}) \end{bmatrix}, \ \forall \mathbf{x} \in \mathcal{X}, \tag{7.1}$$

where $\mathbf{a}_i^T$ is the i^{th} row of the full rank matrix $\mathbf{A} \in \mathbb{R}^{m\times n}(m < n)$, $\{g_i(\cdot)\}_{i=1}^m$ are the functions on $\mathbb{R}$, and $\mathcal{X}$ is the set of signals on which dimensionality reduction is desired. A natural question that arises here is about the circulant approximation to A being a good approximation to non-linear dimensionality reduction operation. Specifically, how *good* is the following approximation

$$f(\mathbf{x}) \approx \begin{bmatrix} g_1(\hat{\mathbf{a}}_1^T \mathbf{x}) \\ \vdots \\ g_m(\hat{\mathbf{a}}_m^T \mathbf{x}) \end{bmatrix} \tag{7.2}$$

where $\hat{\mathbf{a}}_i^T$ is the i^{th} row of the circulant approximation to matrix $\mathbf{A}$.

Another advantage of analyzing the non-linear dimensionality reduction as described above is that we can extend it to understanding of so called *multi-index* function which are defined over unit euclidean ball in $\mathbb{R}^n$ as follows

$$h(\mathbf{x}) = \sum_{i=1}^{m} g_i(\mathbf{a}_i^T \mathbf{x}),\ \mathbf{x} \in \mathcal{X} = \{\mathbf{x} \in \mathbb{R}^n, \|\mathbf{x}\|_2 \leq 1\},$$

Multi-index functions of above form arise in many areas including well known neural networks [173] where $g(u) = \frac{1}{1+e^{-t}}$ is the *sigmoid* function. This analysis can be potentially extended to the understanding of *convolutional neural networks* which are special case of neural networks with parameter sharing and sparse connections. Our results on approximating a given LDR matrix with a circulant matrix might have implications on approximation of a dense neural network with a *convolutional neural network.* These directions of exploration have potential to give novel theoretical insights into the workings of convolutional neural networks.

7.2 Noisy matrix and tensor completion under sparse factor models.

For noisy tensor completion so far we have only explored Gaussian noise density. Similar to [2] the generic error bound for tensor completion can be instantiated for other noise densities. Structured tensor factor models different from sparse CP decomposition could also be explored in future research efforts. In addition to this, an important open question is about the tightness of the error bound for tensor completion. An important direction of future research is towards establishing the optimality of bounds by obtaining *mini-max* lower bounds for noisy tensor completion, similar to ones established for noisy matrix completion in [174]. Finally, developing numerical algorithms with provable convergence to global minima is yet another challenging task that could be explored in future research efforts.

7.3 Matrix completion from noisy and quantized observations

So we have obtained error bounds for the set of low rank matrices or the matrices following the sparse factor models. However, there are many other interesting applications where factor models with different structural constraints on the factors. For example, the non-negative matrix factorization $\mathcal{A} = \left\{\mathbf{A}\middle|\forall a_{ij} \geq 0\right\}$, $\mathcal{D} = \left\{\mathbf{D}\middle|\forall d_{ij} \geq 0\right\}$, arise in text mining, blind source separation and many more areas. Our error bounds can be extended to these sets. Leveraging our approach to gain an understanding of quantized matrix completion for such structured matrix sets could be an interesting future direction to pursue.

Many structured matrix sets are non-convex and accordingly the constrained maximum likelihood approach of matrix completion for such matrix sets involves solving a non-convex problem. Developing algorithms with provable global convergence guarantees is one of the most challenging task. Recently algorithms for projections on the sparse symmetric convex sets which perform better than iterative hard thresholding have been proposed in [175]. Another recent work has proposed an unifying framework for convergence analysis of projected gradient descent algorithm [176]. Several other recent papers have revealed that with proper initialization the non-convex can be solved globally [177–179]. It is an open question whether the techniques outlined in these works can be extended to quantized matrix completion for non-convex structured matrix sets.

References

[1] S. Jain and J. Haupt. Convolutional approximations to linear dimensionality reduction operators. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2017.

[2] A. Soni, S. Jain, J. Haupt, and S. Gonella. Noisy matrix completion under sparse factor models. *IEEE Transactions on Information Theory*, 62(6):3636–3661, 2016.

[3] A. Soni, S. Jain, J. Haupt, and S. Gonella. Error bounds for maximum likelihood matrix completion under sparse factor models. In *2014 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pages 399–403, Dec 2014.

[4] S. Jain, Gutierrez A, and J. Haupt. Noisy tensor completion for tensors with a sparse canonical polyadic factor. In *2017 IEEE International Symposium on Information Theory (ISIT)*, June 2017.

[5] S. Jain, A. Soni, J. Haupt, N. Rao, and R. Nowak. Knowledge-enhanced compressive measurement designs for estimating sparse signals in clutter. *Proc. Signal Processing with Adaptive Sparse Structured Representations (SPARS)*, 2013.

[6] S. Jain, A. Soni, and J. Haupt. Compressive measurement designs for estimating structured signals in structured clutter: A Bayesian experimental design approach. In *IEEE Asilomar Conference on Signals, Systems and Computers*, pages 163–167. IEEE, 2013.

[7] S. Jain, U. Oswal, K. Xu, B. Eriksson, and J. Haupt. A compressed sensing based decomposition of electrodermal activity signals. *IEEE Transactions on Biomedical Engineering*, 2016.

[8] S. Jain, S. J. Kim, and G. B. Giannakis. Backhaul-constrained multicell cooperation leveraging sparsity and spectral clustering. *IEEE Transactions on Wireless Communications*, 15(2):899–912, Feb 2016.

[9] J. Druce, S. Gonella, M. Kadkhodaie, S. Jain, and J. Haupt. Locating material defects via wavefield demixing with morphologically germane dictionaries. *Structural Health Monitoring*, 16(1):112–125, 2017.

[10] S. J. Kim, S. Jain, and G. B. Giannakis. Backhaul-constrained multi-cell cooperation using compressive sensing and spectral clustering. In *Signal Processing Advances in Wireless Communications (SPAWC), 2012 IEEE 13th International Workshop on*, pages 65–69. IEEE, 2012.

[11] M. Kadkhodaie, S. Jain, J. Haupt, J. Druce, and S. Gonella. Locating rare and weak material anomalies by convex demixing of propagating wavefields. In *Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP), 2015 IEEE 6th International Workshop on*, pages 373–376. IEEE, 2015.

[12] M. Kadkhodaie, S. Jain, J. Haupt, J. Druce, and S. Gonella. Group-level support recovery guarantees for group lasso estimation. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2017.

[13] A. Soni, T. Chevalier, and S. Jain. Noisy inductive matrix completion under sparse factor models. In *2017 IEEE International Symposium on Information Theory (ISIT)*, June 2017.

[14] K. Pearson. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, 1901.

[15] I. Jolliffe. *Principal component analysis*. Wiley Online Library, 2002.

[16] J. P. Cunningham and Z. Ghahramani. Unifying linear dimensionality reduction. *arXiv preprint arXiv:1406.0873*, 2014.

[17] Y. C. Eldar and G. Kutyniok. *Compressed Sensing: Theory and Applications*. Cambridge University Press, 2012.

[18] M. Elad. Optimized projections for compressed sensing. *IEEE Transactions on Signal Processing*, 55(12):5695–5702, 2007.

[19] M. Rudelson and R. Vershynin. On sparse reconstruction from Fourier and Gaussian measurements. *Communications on Pure and Applied Mathematics*, 61(8):1025–1045, 2008.

[20] A. Ashok, P. K. Baheti, and M. A. Neifeld. Compressive imaging system design using task-specific information. *Applied Optics*, 47(25):4457–4471, 2008.

[21] J. M. Duarte-Carvajalino and G. Sapiro. Learning to sense sparse signals: Simultaneous sensing matrix and sparsifying dictionary optimization. *IEEE Transactions on Image Processing*, 18(7):1395–1408, 2009.

[22] R. Calderbank, S. Howard, and S. Jafarpour. Construction of a large class of deterministic sensing matrices that satisfy a statistical isometry property. *IEEE Journal of Selected Topics in Signal Processing*, 4(2):358–374, 2010.

[23] J. Haupt, L. Applebaum, and R. Nowak. On the restricted isometry of deterministically subsampled Fourier matrices. In *IEEE Conference on Information Sciences and Systems*, pages 1–6, 2010.

[24] L. Zelnik-Manor, K. Rosenblum, and Y. C. Eldar. Sensing matrix optimization for block-sparse decoding. *IEEE Transactions on Signal Processing*, 59(9):4300–4312, 2011.

[25] A. Ashok and M. A. Neifeld. Compressive imaging: hybrid measurement basis design. *Journal of the Optical Society of America A*, 28(6):1041–1050, 2011.

[26] W. R. Carson, M. Chen, M. R. D. Rodrigues, R. Calderbank, and L. Carin. Communications-inspired projection design with application to compressive sensing. *SIAM Journal on Imaging Sciences*, 5(4):1185–1212, 2012.

[27] S. Ji, Y. Xue, and L. Carin. Bayesian compressive sensing. *IEEE Transactions on Signal Processing*, 56(6):2346–2356, 2008.

[28] J. Haupt, R. Baraniuk, R. Castro, and R. Nowak. Compressive distilled sensing: Sparse recovery using adaptivity in compressive measurements. In *Asilomar Conference on Signals, Systems, and Computers*, pages 1551–1555, 2009.

[29] J. Haupt, R. Castro, and R. Nowak. Adaptive sensing for sparse signal recovery. In *Proceedings of IEEE DSP Workshop and Workshop on Signal Processing Education*, pages 702–707, 2009.

[30] J. Haupt and R. Nowak. Adaptive sensing for sparse recovery. In Y. Eldar and G. Kutyniok, editors, *Compressed Sensing: Theory and Applications*. Cambridge University Press, 2011.

[31] P. Indyk, E. Price, and D. P. Woodruff. On the power of adaptivity in sparse recovery. In *Proceedings of IEEE Foundations of Computer Science*, pages 285–294, 2011.

[32] A. Ashok, J. L. Huang, and M. A. Neifeld. Information-optimal adaptive compressive imaging. In *Asilomar Conference on Signals, Systems and Computers*, pages 1255–1259, 2011.

[33] M. Iwen and A. Tewfik. Adaptive group testing strategies for target detection and localization in noisy environments. *IEEE Transactions on Signal Processing*, 60(5):2344–2353, 2012.

[34] J. Haupt, R. Baraniuk, R. Castro, and R. Nowak. Sequentially designed compressed sensing. In *Proceedings of IEEE Statistical Signal Processing Workshop*, pages 401–404, 2012.

[35] M. A. Davenport and E. Arias-Castro. Compressive binary search. In *IEEE International Symposium on Information Theory (ISIT)*, pages 1827–1831, 2012.

[36] S. Balakrishnan, M. Kolar, A. Rinaldo, and A. Singh. Recovering block-structured activations using compressive measurements. *arXiv preprint arXiv:1209.3431*, 2012.

[37] J. M. Duarte-Carvajalino, G. Yu, L. Carin, and G. Sapiro. Task-driven adaptive statistical compressive sensing of gaussian mixture models. *IEEE Transactions on Signal Processing*, 61(3):585–600, 2013.

[38] J. Sharpnack A. Krishnamuthy and A. Singh. Recovering graph-structured activations using adaptive compressive measurements. In *Proceedings of Asilomar Conference on Signals, Systems and Computers*, pages 765–769, 2013.

[39] E. Tánczos and R. M. Castro. Adaptive sensing for estimation of structured sparse signals. *arXiv preprint arXiv:1311.7118*, 2013.

[40] A. Soni and J. Haupt. On the fundamental limits of recovering tree sparse vectors from noisy linear measurements. *IEEE Transactions on Information Theory*, 60(1):133–149, 2014.

[41] M. Malloy and R. Nowak. Near-optimal adaptive compressive sensing. *IEEE Transactions on Information Theory*, 60(4):4001–4012, 2014.

[42] R. M. Castro. Adaptive sensing performance lower bounds for sparse signal detection and support estimation. *Bernoulli*, 20(4):2217–2246, 2014.

[43] R. M. Castro and E. Tánczos. Adaptive compressed sensing for estimation of structured sparse sets. *arXiv preprint arXiv:1410.4593*, 2014.

[44] G. Strang. A proposal for Toeplitz matrix calculations. *Studies in Applied Mathematics*, 74(2):171–176, 1986.

[45] T. F. Chan. An optimal circulant preconditioner for Toeplitz systems. *SIAM Journal on Scientific and Statistical Computing*, 9(4):766–771, 1988.

[46] E. E. Tyrtyshnikov. Optimal and superoptimal circulant preconditioners. *SIAM Journal on Matrix Analysis and Applications*, 13(2):459–473, 1992.

[47] J. Müller-Quade, H. Aagedal, T. Beth, and M. Schmid. Algorithmic design of diffractive optical systems for information processing. *Physica D: Nonlinear Phenomena*, 120(1):196–205, 1998.

[48] M. Schmid, R. Steinwandt, J. Müller-Quade, M. Rötteler, and T. Beth. Decomposing a matrix into circulant and diagonal factors. *Linear Algebra and its Applications*, 306(1):131–143, 2000.

[49] M. Huhtanen. How real is your matrix? *Linear algebra and its applications*, 424(1):304–319, 2007.

[50] M. Huhtanen. Factoring matrices into the product of two matrices. *BIT Numerical Mathematics*, 47(4):793–808, 2007.

[51] M. Huhtanen. Approximating ideal diffractive optical systems. *Journal of Math. Analysis and Applications*, 345(1):53–62, 2008.

[52] J. Romberg. Compressive sensing by random convolution. *SIAM Journal on Imaging Sciences*, 2(4):1098–1128, 2009.

[53] J. Haupt, W. U. Bajwa, G. Raz, and R. Nowak. Toeplitz compressed sensing matrices with applications to sparse channel estimation. *IEEE Transactions on Information Theory*, 56(11):5862–5875, 2010.

[54] H. Rauhut. Compressive sensing and structured random matrices. In M. Fornasier, editor, *Theoretical Foundations and Numerical Methods for Sparse Recovery*, volume 9, pages 1–92. deGruyter, 2010.

[55] H. Rauhut, J. Romberg, and J. A. Tropp. Restricted isometries for partial random circulant matrices. *Applied and Computational Harmonic Analysis*, 32(2):242–254, 2012.

[56] F. Krahmer and R. Ward. New and improved Johnson-Lindenstrauss embeddings via the restricted isometry property. *SIAM Journal on Mathematical Analysis*, 43(3):1269–1281, 2011.

[57] H. L. Yap, M. B. Wakin, and C. J. Rozell. Stable manifold embeddings with structured random matrices. *IEEE Journal of Selected Topics in Signal Processing*, 7(4):720–730, 2013.

[58] H. Zhang and L. Cheng. New bounds for circulant Johnson-Lindenstrauss embeddings. *arXiv preprint arXiv:1308.6339*, 2013.

[59] N. Parikh and S. Boyd. Proximal algorithms. *Foundations and Trends® in Optimization*, 1(3):127–239, 2014.

[60] J. Liu, S. Ji, and J. Ye. *SLEP: Sparse Learning with Efficient Projections*. Arizona State University, 2009.

[61] P. Jain, P. Netrapalli, and S. Sanghavi. Low-rank matrix completion using alternating minimization. In *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*, pages 665–674. ACM, 2013.

[62] N. Ailon and B. Chazelle. Approximate nearest neighbors and the fast Johnson-Llindenstrauss transform. In *Proceedings of ACM Symposium on Theory of Computing*, pages 557–563, 2006.

[63] A. Dasgupta, R. Kumar, and T. Sarlós. A sparse Johnson-Lindenstrauss transform. In *Proceedings of ACM Symposium on Theory of Computing*, pages 341–350, 2010.

[64] J. Hopcroft and R. Kannan. *Foundations of Data Science*. April 2014. Available online: `http://www.cs.cornell.edu/jeh/book11April2014.pdf`.

[65] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge university press, 2004.

[66] V. Chandrasekaran, B. Recht, P. Parrilo, and A. Willsky. The convex geometry of linear inverse problems. *Foundations of Computational mathematics*, 12(6):805–849, 2012.

[67] D. Klain and G. Rota. *Introduction to geometric probability*. Cambridge University Press, 1997.

[68] M. A. Davenport, P. T. Boufounos, M. B. Wakin, and R. G. Baraniuk. Signal processing with compressive measurements. *IEEE Journal of Selected Topics in Signal Processing*, 4(2):445–460, 2010.

[69] R. Burkard and U. Derigs. The linear sum assignment problem. In *Assignment and Matching Problems: Solution Methods with FORTRAN-Programs*, pages 1–15. Springer, 1980.

[70] E. J. Candès and B. Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9(6):717–772, 2009.

[71] E. J. Candès and T. Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE Transactions on Information Theory*, 56(5):2053–2080, 2010.

[72] R. H. Keshavan, A. Montanari, and S. Oh. Matrix completion from a few entries. *IEEE Transactions on Information Theory*, 56(6):2980–2998, 2010.

[73] B. Recht. A simpler approach to matrix completion. *The Journal of Machine Learning Research*, 12:3413–3430, 2011.

[74] E. J. Candes and Y. Plan. Matrix completion with noise. *Proceedings of the IEEE*, 98(6):925–936, 2010.

[75] R. Keshavan, A. Montanari, and S. Oh. Matrix completion from noisy entries. In *Advances in Neural Information Processing Systems*, pages 952–960, 2009.

[76] V. Koltchinskii, K. Lounici, and A. B.s Tsybakov. Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *The Annals of Statistics*, 39(5):2302–2329, 2011.

[77] Y. Koren, R. Bell, and C. Volinsky. Matrix factorization techniques for recommender systems. *Computer*, (8):30–37, 2009.

[78] J. D. M. Rennie and N. Srebro. Fast maximum margin matrix factorization for collaborative prediction. In *Proceedings of the 22nd international conference on Machine learning*, pages 713–719. ACM, 2005.

[79] A. S. Lan, C. Studer, and R. G. Baraniuk. Matrix recovery from quantized and corrupted measurements. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pages 4973–4977. IEEE, 2014.

[80] A. Karbasi and S. Oh. Robust localization from incomplete local information. *Networking, IEEE/ACM Transactions on*, 21(4):1131–1144, 2013.

[81] T. G. Kolda and B. W. Bader. Tensor decompositions and applications. *SIAM review*, 51(3):455–500, 2009.

[82] N. D. Sidiropoulos, L. De Lathauwer, X. Fu, K. Huang, E. E. Papalexakis, and C. Faloutsos. Tensor decomposition for signal processing and machine learning. *arXiv preprint arXiv:1607.01668*, 2016.

[83] M. Yuan and C. Zhang. On tensor completion via nuclear norm minimization. *Foundations of Computational Mathematics*, 16(4):1031–1068, 2016.

[84] J. Liu, P. Musialski, P. Wonka, and J. Ye. Tensor completion for estimating missing values in visual data. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(1):208–220, 2013.

[85] B. Huang, C. Mu, D. Goldfarb, and J. Wright. Provable low-rank tensor recovery. *Optimization-Online*, 4252:2, 2014.

[86] E. Elhamifar and R. Vidal. Sparse subspace clustering: Algorithm, theory, and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(11):2765–2781, 2013.

[87] B. Eriksson, L. Balzano, and R. Nowak. High-rank matrix completion and subspace clustering with missing data. *arXiv preprint arXiv:1112.5629*, 2011.

[88] M. Soltanolkotabi and E. J. Candès. A geometric analysis of subspace clustering with outliers. *The Annals of Statistics*, 40(4):2195–2238, 2012.

[89] M. Aharon, M. Elad, and A. Bruckstein. K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Transactions on Signal Processing*, 54(11):4311–4322, 2006.

[90] J. Mairal, M. Elad, and G. Sapiro. Sparse representation for color image restoration. *IEEE Transactions on Image Processing*, 17(1):53–69, 2008.

[91] B. A. Olshausen and D. J. Field. Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision Research*, 37:3311–3325, 1997.

[92] G. Allen. Sparse higher-order principal components analysis. In *AISTATS*, volume 15, 2012.

[93] R. Ruiters and R. Klein. BTF compression via sparse tensor decomposition. In *Computer Graphics Forum*, volume 28, pages 1181–1188. Wiley Online Library, 2009.

[94] W. Shi, Y. Zhu, S. Y. Philip, M. Liu, G. Wang, Z. Qian, and Y. Lian. Incomplete electrocardiogram time series prediction. In *Biomedical Circuits and Systems Conference (BioCAS), 2016 IEEE*, pages 200–203. IEEE, 2016.

[95] E. E. Papalexakis, N. D. Sidiropoulos, and R. Bro. From k-means to higher-way co-clustering: Multilinear decomposition with sparse latent factors. *IEEE transactions on signal processing*, 61(2):493–506, 2013.

[96] Y. Pang, Z. Ma, J. Pan, and Y. Yuan. Robust sparse tensor decomposition by probabilistic latent semantic analysis. In *Sixth International Conference on Image and Graphics (ICIG)*, pages 893–896. IEEE, 2011.

[97] N. Srebro, N. Alon, and T. S. Jaakkola. Generalization error bounds for collaborative prediction with low-rank matrices. In *Advances In Neural Information Processing Systems*, pages 1321–1328, 2004.

[98] M. A. Davenport, Y. Plan, E. van den Berg, and M. Wootters. 1-bit matrix completion. *Information and Inference*, 3(3):189–223, 2014.

[99] Y. Plan, R. Vershynin, and E. Yudovina. High-dimensional estimation with geometric constraints. *arXiv preprint arXiv:1404.3749*, 2014.

[100] A. V. Sambasivan and J. Haupt. Minimax lower bounds for noisy matrix completion under sparse factor models. *arXiv preprint arXiv:1510.00701*, 2015.

[101] Y. Xu, R. Hao, W. Yin, and Z. Su. Parallel matrix factorization for low-rank tensor completion. *arXiv preprint arXiv:1312.1254*, 2013.

[102] P. Jain and S. Oh. Provable tensor factorization with missing data. In *Advances in Neural Information Processing Systems*, pages 1431–1439, 2014.

[103] T. M. Cover and J. A. Thomas. *Elements of information theory*. John Wiley & Sons, 2012.

[104] T. Cai and W. Zhou. A max-norm constrained minimization approach to 1-bit matrix completion. *The Journal of Machine Learning Research*, 14(1):3619–3647, 2013.

[105] Y. Cao and Y. Xie. Categorical matrix completion. In *Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP), 2015 IEEE 6th International Workshop on*, pages 369–372. IEEE, 2015.

[106] O. Klopp, J. Lafond, É. Moulines, and J. Salmon. Adaptive multinomial matrix completion. *Electronic Journal of Statistics*, 9(2):2950–2975, 2015.

[107] S. A. Bhaskar. Probabilistic low-rank matrix completion from quantized measurements. *Journal of Machine Learning Research*, 17(60):1–34, 2016.

[108] J. Haupt, N. Sidiropoulos, and G. B. Giannakis. Sparse dictionary learning from 1-bit data. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7664–7668. IEEE, 2014.

[109] Y. Shen, M. Mardani, and G. B. Giannakis. Online categorical subspace learning for sketching big data with misses. *arXiv preprint arXiv:1609.08235*, 2016.

[110] Y. Shen and G. B. Giannakis. Online dictionary learning from large-scale binary data. In *2016 24th European Signal Processing Conference (EUSIPCO)*, pages 1808–1812, Aug 2016.

[111] M. Talagrand. *The generic chaining: upper and lower bounds of stochastic processes*. Springer Science & Business Media, 2006.

[112] M. Talagrand. Majorizing measures: the generic chaining. *The Annals of Probability*, pages 1049–1103, 1996.

[113] A. Zymnis, S. Boyd, and E. Candes. Compressed sensing with quantized measurements. *IEEE Signal Processing Letters*, 17(2):149–152, Feb 2010.

[114] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning*, 3(1):1–122, 2011.

[115] T. Blumensath and M. E. Davies. Iterative hard thresholding for compressed sensing. *Applied and computational harmonic analysis*, 27(3):265–274, 2009.

[116] C. C. Craig. On the Tchebychef inequality of Bernstein. *The Annals of Mathematical Statistics*, 4(2):94–102, 1933.

[117] S. Dirksen. Tail bounds via generic chaining. *Electronic Journal of Probability*, 20, 2015.

[118] M. Rudelson and R. Vershynin. Non-asymptotic theory of random matrices: extreme singular values. *arXiv preprint arXiv:1003.2990*, 2010.

[119] E. J. Candès and T. Tao. The Dantzig selector: Sstatistical estimation when p is much larger than n. *Ann. Statist*, 35:2313–2351, 2007.

[120] R. Baraniuk, M. Davenport, R. DeVore, and M. Wakin. A simple proof of the restricted isometry property for random matrices. *Constructive Approximation*, 28(3):253–263, 2008.

[121] J. Huang, T. Zhang, and D. Metaxas. Learning with structured sparsity. *Journal of Machine Learning Research*, 12:3371–3412, 2011.

[122] R. Baraniuk, V. Cevher, M. Duarte, and C Hegde. Model-based compressive sensing. *IEEE Transactions on Information Theory*, 56(4):1982–2001, 2010.

[123] M. Elad. Optimized projections for compressed sensing. *IEEE Transactions on Signal Processing*, 55:5695–5702, 2007.

[124] J. M. Duarte-Carvajalino and G. Sapiro. Learning to sense sparse signals: Simultaneous sensing matrix and sparsifying dictionary optimization. *IEEE Transactions on Image Processing*, 18:1395–1408, 2009.

[125] K. Rosenblum, L. Zelnik-Manor, and Y. C. Eldar. Sensing matrix optimization for block-sparse decoding. *IEEE Transactions on Signal Processing*, 59(9):4300–4312, 2011.

[126] G. Yu J. M. Duarte-Carvajalino, L. Carin, and G. Sapiro. Task-driven adaptive statistical compressive sensing of gaussian mixture models. *IEEE Transactions on Signal Processing*, 61(3):585–600, 2013.

[127] N. Rao and R. Nowak. Correlated Gaussian designs for compressive imaging. In *Proceedings of IEEE International Conference on Image Processing*, 2012.

[128] G. Reeves and M. Gastpar. Differences between observation and sampling error in sparse signal reconstruction. In *Statistical Signal Processing, 2007. SSP'07. IEEE/SP 14th Workshop on*, pages 690–694. IEEE, 2007.

[129] S. Aeron, V. Saligrama, and M. Zha. Information theoretic bounds for compressed sensing. *IEEE Transactions on Information Theory*, 56:5111–5130, 2010.

[130] E. Arias-Castro and Y. Eldar. Noise folding in compressed sensing. *IEEE Signal Processing Letters*, 18:478–481, 2011.

[131] K. Krishnamurthy, R. Willett, and M. Raginsky. Target detection performance bounds in compressive imaging. *EURASIP Journal on Advances in Signal Processing*, 2012.

[132] S. Ji, Y. Xue, and L. Carin. Bayesian compressive sensing. *IEEE Transactions on Signal Processing*, 56(6):2346–2356, 2008.

[133] M. W. Seeger. Bayesian inference and optimal design for the sparse linear model. *The Journal of Machine Learning Research*, 9:759–813, 2008.

[134] M. W. Seeger and H. Nickisch. Compressed sensing and Bayesian experimental design. In *Proceedings of International Conference on Machine Learning*, pages 912–919, 2008.

[135] M. Seeger and H. Nickisch. Large scale Bayesian inference and experimental design for sparse linear models. *SIAM Journal of Imaging Sciences*, 4(1):166–199, 2011.

[136] P. Schniter. Exploiting structured sparsity in Bayesian experimental design. In *Proceedings of IEEE Workshop on Computational Advances in Multi-Sensor Adaptive Processing*, pages 357–360, 2011.

[137] M. Seeger, H. Nickisch, R. Pohmann, and B. Schölkopf. Optimization of k-space trajectories for compressed sensing by Bayesian experimental design. *Magnetic resonance in medicine*, 63(1):116–126, 2010.

[138] W. R. Carson, M. R. D. Rodrigues, M. Chen, L. Carin, and R. Calderbank. How to focus the discriminative power of a dictionary. In *Proc. ICASSP*, pages 1365–1368, 2012.

[139] H. L. Van Trees. *Detection, Estimation, and Modulation Theory, Optimum Array Processing.* Wiley-Interscience, 2004.

[140] A. Scaglione, P. Stoica, S. Barbarossa, G. B. Giannakis, and H. Sampath. Optimal designs for space-time linear precoders and decoders. *IEEE Transactions on Signal Processing*, 50(5):1051–1064, 2002.

[141] S. M. Kay. *Fundamentals of statistical signal processing, Volume 1: Estimation theory.* Prentice Hall PTR, 1993.

[142] I. D. Schizas, G. B. Giannakis, and Z.-Q. Luo. Distributed estimation using reduced-dimensionality sensor observations. *IEEE Transactions on Signal Processing*, 55(8):4284–4299, 2007.

[143] S. Jain, A. Soni, J. Haupt, N. Rao, and R. Nowak. Knowledge-enhanced compressive measurement designs for estimating sparse signals in clutter. In *Proc. Signal Processing with Adaptive Sparse Structured Representations (SPARS)*, 2013.

[144] L. L. Scharf. *Statistical signal processing: Detection, Estimation, and Time Series Analysis.* Addison-Wesley, Reading, MA, 1991.

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

[146] D. P. Palomar and S. Verdú. Gradient of mutual information in linear vector Gaussian channels. *IEEE Transactions on Information Theory*, 52(1):141–154, 2006.

[147] N. Taylor and C. Machado-Moreira. Regional variations in transepidermal water loss, eccrine sweat gland density, sweat secretion rates and electrolyte composition in resting and exercising humans. *Extreme physiology & medicine*, 2:4, 2013.

[148] T. Nishiyama et al. Irregular activation of individual sweat glands in human sole observed by a videomicroscopy. *Autonomic Neuroscience*, 88:117–126, 2001.

[149] B. Sidis. The nature and cause of the galvanic phenomenon. *The Journal of Abnormal Psychology*, 5(2):69–74, 1910.

[150] Liberate yourself from the lab: Q Sensor measures EDA in the wild. White paper, Affectiva Inc., 2012.

[151] M. Garbarino et al. Empatica e3 - a wearable wireless multi-sensor device for real-time computerized biofeedback and data acquisition. In *2014 EAI 4th International Conference on Wireless Mobile Communication and Healthcare (Mobihealth)*, pages 39–42, 2014.

[152] M. Benedek, and C. Kaernbach. Decomposition of skin conductance data by means of nonnegative deconvolution. *Psychophysiology*, 47:647–658, 2010.

[153] F. Silveira, B. Eriksson, A. Sheth, and A. Sheppard. Predicting audience responses to movie content from electro-dermal activity signals. In *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing*, pages 707–716. ACM, 2013.

[154] W. Lian et al. Modeling correlated arrival events with latent semi-Markov processes. In *Proceedings of 31st International Conference of Machine Learning*, pages 396–404, 2014.

[155] H. Lu et al. StressSense: Detecting stress in unconstrained acoustic environments using smartphones. In *Proceedings of ACM Conference Ubiquitous Computing*, pages 351–360, 2012.

[156] C. L. Lim, C. Rennie, R. J. Barry, H. Bahramali, I. Lazzaro, B. Manor, and E. Gordon. Decomposing skin conductance into tonic and phasic components. *International Journal of Psychophysiology*, 25(2):97–109, 1997.

[157] D. M. Alexander, C. Trengove, P. Johnston, T. Cooper, JP August, and E. Gordon. Separating individual skin conductance responses in a short interstimulus-interval paradigm. *Journal of neuroscience methods*, 146(1):116–123, 2005.

[158] D. Bach et al. Dynamic causal modeling of spontenous fluctuations in skin conductance. *Psychophysiology*, 48:1–6, 2010.

[159] A. Greco, G. Valenza, Antonio L., E. P. Scilingo, and L. Citi. cvxeda: A convex optimization approach to electrodermal activity processing. *IEEE Transactions on Biomedical Engineering*, 63(4):797–804, 2016.

[160] T. Chaspari, A. Tsiartas, L. I. Stein, S. A. Cermak, and S. S. Narayanan. Sparse representation of electrodermal activity with knowledge-driven dictionaries. *IEEE Transactions on Biomedical Engineering*, 62(3):960–971, 2015.

[161] M. B. McCoy and J. A. Tropp. The achievable performance of convex demixing. *arXiv preprint arXiv:1309.7478 [cs.IT]*, 2013.

[162] M. B. McCoy and J. A. Tropp. Sharp recovery bounds for convex demixing, with applications. *Foundations of Computational Mathematics*, 14(3):503–567, 2014.

[163] D. L. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306, 2006.

[164] W. Yin, S. Morgan, J. Yang, and Y. Zhang. Practical compressive sensing with Toeplitz and circulant matrices. In *Visual Communications and Image Processing 2010*, page 77440K. International Society for Optics and Photonics, 2010.

[165] C. R. Berger, S. Zhou, J. C. Preisig, and P. Willett. Sparse channel estimation for multicarrier underwater acoustic communication: From subspace methods to compressed sensing. *IEEE Transactions on Signal Processing*, 58(3):1708–1721, 2010.

[166] H. Critchley et al. Neural activity relating to generation and representation of galvanic skin conductance responses: A functional magnetic resonance imaging study. *Journal of Neuroscience*, 20(8):3033–3040, 2000.

[167] J. Healey et al. Out of the lab and into the fray: Towards modeling emotion in everyday life. In *International Conference on Pervasive Computing*, pages 156–173, 2010.

[168] R. Foygel and L. Mackey. Corrupted sensing: Novel guarantees for separating structured signals. *IEEE Transactions on Information Theory*, 60(2):1223–1247, 2014.

[169] C. Studer and R. G. Baraniuk. Stable restoration and separation of approximately sparse signals. *Applied and Computational Harmonic Analysis*, 37(1):12–35, 2014.

[170] M. Grant and S. Boyd. CVX: Matlab software for disciplined convex programming, version 2.1, March 2014.

[171] S. Diamond and S. Boyd. Matrix-free convex optimization modeling. *arXiv preprint arXiv:1506.00760 [math.OC]*, 2015.

[172] S. R. Becker, E. J. Candès, and M. C. Grant. TFOCS: Templates for first-order conic solvers, 2012.

[173] A. Pinkus. Approximation theory of the mlp model in neural networks. *Acta Numerica*, 8:143–195, 1999.

[174] A. V. Sambasivan. Minimax lower bounds for noisy matrix completion under sparse factor models. 2015. online at: `http://www.arxiv.org/abs/1510.00701`.

[175] A. Beck and N. Hallak. On the minimization over sparse symmetric sets. Technical report, Technical report, Technion, 2014.

[176] S. Oymak, B. Recht, and M. Soltanolkotabi. Sharp time–data tradeoffs for linear inverse problems. *arXiv preprint arXiv:1507.04793*, 2015.

[177] A. Agarwal, A. Anandkumar, P. Jain, and P. Netrapalli. Learning sparsely used overcomplete dictionaries via alternating minimization. *arXiv preprint arXiv:1310.7991*, 2013.

[178] S. Arora, R. Ge, and A. Moitra. New algorithms for learning incoherent and overcomplete dictionaries. *arXiv preprint arXiv:1308.6273*, 2013.

[179] S. Arora, R. Ge, T. Ma, and A. Moitra. Simple, efficient, and neural algorithms for sparse coding. *arXiv preprint arXiv:1503.00778*, 2015.