ROBUST LEARNING WITH LOW-DIMENSIONAL STRUCTURES: THEORY, ALGORITHMS AND APPLICATIONS

Yuxiang Wang

B.Eng.(Hons), National University of Singapore

In partial fulfilment of the requirements for the degree of

MASTER OF ENGINEERING

Department of Electrical and Computer Engineering

National University of Singapore

August 2013

DECLARATION

I hereby declare that this thesis is my original work and it has been written by me in its entirety.

I have duly acknowledged all the sources of information which have been used in the thesis.

This thesis has also not been submitted for any degree in any university previously.

Yuxiang Wang

October 25, 2013

Acknowledgements

I would like to thank my research advisors, Prof. Loong-Fah Cheong and Prof. Huan Xu, for their guidance, timely discussions and constant encouragement during my candidature. The key ideas in this thesis could not have emerged and then rigorously formalized without the their profound insights, sharp intuition and technical guidance in every stage of my research. I would also like to thank my collaborators, Prof. Chenlei Leng from the Department of Statistics and Prof. Kim-Chuan Toh from the Department of Mathematics, for their valuable advice in statistics and optimization.

I owe my deep appreciation to my friend Ju Sun, from whom I learned the true meaning of research and scholarship. He was also the one that introduced me to the computer vision and machine learning research two years ago, which I stayed passionate about ever since.

Special thanks to my friends and peer researchers Choon Meng, Chengyao, Jiashi, Xia Wei, Gao Zhi, Zhuwen, Jiaming, Shazor, Lin Min, Lile, Tianfei, Bichao, Zhao Ke and etc for the seminar classes, journal clubs, lunches, dinners, games, pizza parties and all the fun together. Kudos to the our camaraderie!

Finally, I would like to thank my parents for their unconditional love and support during my graduate study, and to my wife Su, for being the amazing delight for me every day.

Contents

Summary

High dimensionality is often considered a "curse" for machine learning algorithms, in a sense that the required amount of data to learn a generic model increases exponentially with dimension. Fortunately, most real problems possess certain low-dimensional structures which can be exploited to gain statistical and computational tractability. The key research question is "How". Since low-dimensional structures are often highly non-convex or combinatorial, it is often NP-hard to impose such constraints. Recent development in compressive sensing and matrix completion/recovery has suggested a way. By combining the ideas in optimization (in particular convex optimization theory), statistical learning theory and high-dimensional geometry, it is sometimes possible to learn these structures exactly by solving a convex surrogate of the original problem. This approach has led to notable advances and in quite a few disciplines such as signal processing, computer vision, machine learning and data mining. Nevertheless, when the data are noisy or when the assumed structures are only a good approximation, learning the parameters of a given structure becomes a much more difficult task.

In this thesis, we study the robust learning of low-dimensional structures when there are uncertainties in the data. In particular, we consider two structures that are common in real problems: "low-rank subspace model" that underlies matrix completion and Robust PCA, and "union-of-subspace model" that arises in the problem of subspace clustering. In the upcoming chapters, we will present (i) stability of matrix factorization and its consequences in the robustness of collaborative filtering (movie recommendations) against manipulators; (ii) sparse subspace clustering under random and deterministic

noise; (iii) simultaneous low-rank and sparse regularization for subspace clustering; and (iv) Proximal Alternating Robust Subspace Minimization (PARSuMi), a robust matrix recovery algorithm that handles simultaneous noise, missing data and gross corruptions. The results in these chapters either solve a real engineering problem or provide interesting insights into why certain empirically strong algorithms succeed in practice. While in each chapter, only one or two real applications are described and demonstrated, the ideas and techniques in this thesis are general and applicable to any problems having the assumed structures.

List of Publications

[1] Y.-X. Wang and H. Xu. Stability of matrix factorization for collaborative filtering. In J. Langford and J. Pineau, editors, *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, ICML '12, pages 417–424, July 2012.

[2] Y.-X. Wang and H. Xu. Noisy sparse subspace clustering. In S. Dasgupta and D. Mcallester, editors, *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, volume 28, pages 89–97. JMLR Workshop and Conference Proceedings, 2013.

[3] Y.-X. Wang, C. M. Lee, L.-F. Cheong, and K. C. Toh. Practical matrix completion and corruption recovery using proximal alternating robust subspace minimization. *Under review for publication at IJCV*, 2013.

[4] Y.-X. Wang, H. Xu, and C. Leng. Provable subspace clustering: When LRR meets SSC. *To appear at Neural Information Processing Systems (NIPS-13)*, 2013.

List of Tables

List of Figures

List of Abbreviations

ADMM	Alternating Direction Methods of Multipliers
ALS	Alternating Least Squares
APG	Accelerated Proximal Gradient
BCD	Block Coordinate Descent
CDF	Cumulative Density Function
CF	Collaborative filtering
fMRI	Functional Magnetic resonance imaging
GiniIndex	Gini Index: a smooth measure of sparsity
iid	Identically and independently distributed
LASSO	Least Absolute Shrinkage and Selection Operator
LDA	Linear Discriminant Analysis
LM	Levenberg-Marquadt
LP	Linear Programming
LRR	Low Rank Representation
LRSSC	Low Rank Sparse Subspace Clustering
MC	Matrix Completion
MF	Matrix Factorization
NLCG	Nonlinear conjugate gradient
NN/kNN	Nearest Neighbour/K Nearest Neighbour
PARSuMi	Proximal Alternating Robust Subspace Minimization

PCA	Principal Component Analysis
PDF	Probability Density Function
QP	Quadratic Programming
RelViolation	Relative Violation: a soft measure of SEP
RIP	Restricted Isometry Property
RMSE	Root Mean Square Error
RPCA	Robust Principal Component Analysis
SDP	Semidefinite Programming
SEP	Self-Expressiveness Property
SfM/NRSfM	Structure from motion/Non-Rigid Structure from Motion
SSC	Sparse Subspace Clustering
SVD	Singular Value Decomposition

Chapter 1

Introduction

We live in the Big Data Era. According to Google CEO Eric Schmidt, the amount of information we create in 2 days in 2010 is the same as we did from the dawn of civilization to 2003 [120][1]. On Facebook alone, there are 1.2 billion users who generate/share 70 billion contents every month in 2012[128]. Among these, 7.5 billion updates are photos [72]. Since a single digital image of modest quality contains more than a million pixels, a routine task of indexing these photos in their raw form involves dealing with a million by billion data matrix. If we consider instead the problem of recommending these photos to roughly 850 million daily active users [72] based on the "likes" and friendship connections, then we are dealing with a billion by billion rating matrix. These data matrices are massive in both size and dimension and are considered impossible to analyze using classic techniques in statistics[48]. The fundamental limitation in the high dimensional statistical estimation is that the number of data points required to successfully fit a general Lipschitz function increases exponentially with the dimension of the data [48]. This is often described metaphorically as the "curse of dimensionality".

Similar high dimensional data appear naturally in many other engineering problems too, e.g., image/video segmentation and recognition in computer vision, fMRI in medical image processing and DNA microarray analysis in bioinformatics. The data are even more ill-posed in these problems as the dimension is typically much larger than number

[1] That's approx. 5×10^{21} binary bit of data according to the reference.

of data points, making it hard to fit even a linear regression model. The prevalence of such data in real applications makes it a fundamental challenge to develop techniques to better harness the high dimensionality.

The key to overcome this "curse of dimensionality" is to identify and exploit the underlying structures in the data. Early examples of this approach include principal component analysis (PCA)[78] that selects an optimal low-rank approximation in the ℓ_2 sense and linear discriminant analysis (LDA)[88] that maximizes class discrimination for categorical data. Recent development has further revealed that when the data indeed obey certain low-dimensional structures, such as sparsity and low-rank, the high dimensionality can result in desirable data redundancy which makes it possible to *provably* and *exactly* recover the correct parameters of the structure by solving a convex relaxation of the original problem, even when data are largely missing (e.g., matrix completion [24]) and/or contaminated with gross corruptions (e.g., LP decoding [28] and robust PCA [27]). This amazing phenomenon is often referred to as the "blessing of dimensionality"[48].

One notable drawback of these convex optimization-based approaches is that they typically require the data to exactly follow the given structure, namely free of noise and model uncertainties. Real data, however, are at best well-approximated by the structure. Noise is ubiquitous and there are sometimes adversaries intending to manipulate the system to the worst possible. This makes robustness, i.e. the resilience to noise/uncertainty, a desideratum in any algorithm design.

Robust extensions of the convex relaxation methods do exist for sparse and low-rank structures (see [49][22][155]), but their stability guarantees are usually weak and their empirical performances are often deemed unsatisfactory for many real problems (see our discussions and experiments in Chapter 5). Furthermore, when the underlying dimension is known in prior, there is no intuitive way to restrict the solution to be of the desirable dimension as one may naturally do in classic PCA. Quite on the contrary, rank-constrained methods such as matrix factorization are widely adopted in practice but, perhaps due to its non-convex formulation, lack of proper theoretical justification.

For other structures, such as the union-of-subspace model, provable robustness is still an open problem.

This thesis focuses on understanding and developing methodology in the robust learning of low-dimensional structures. We contribute to the field by providing both theoretical analysis and practically working algorithms to robustly learn the parameterization of two important types of models: **low-rank subspace model** and the **union-of-subspace model**. For the prior, we developed the first stability bound for matrix factorization with missing data with applications to the robustness of recommendation systems against manipulators. On the algorithmic front, we derived PARSuMi, a robust matrix completion algorithm with explicit rank and cardinality constraints that demonstrates superior performance in real applications such as structure from motion and photometric stereo. For the latter, we proposed an algorithm called Lasso-SSC that can obtain provably correct subspace clustering even when data are noisy (the first of its kind). We also proposed and analyzed the performance of LRSSC, a new method that combines the advantages of two state-of-the-art algorithms. The results reveal an interesting tradeoff between two performance metrics in the subspace clustering problem.

It is important to note that while our treatments of these problems are mainly theoretical, there are always clear real problems in computer vision and machine learning that motivate the analysis and we will relate to the motivating applications throughout the thesis.

1.1 Low-Rank Subspace Model and Matrix Factorization

Ever since the advent of compressive sensing[50][30][28], the use of ℓ_1 norm to promote sparsity has received tremendous attention. It is now well-known that a sparse signal can be perfectly reconstructed from a much smaller number of samples than what Nyquist-Shannon sampling theorem requires via ℓ_1 norm minimization if the measurements are taken with a sensing matrix that obeys the the so-called restricted isometry property (RIP) [50][20]. This result can also be equivalently stated as correcting sparse

errors in a decoding setting [28] or as recovering a highly incomplete signal in the context of signal recovery[30].

In computer vision, sparse representation with overcomplete dictionaries leads to breakthroughs in image compression[1], image denoising[52], face recognition[148], action/activity recognition[33] and many other problems. In machine learning, it brings about advances and new understanding in classification [74], regression [85], clustering [53] and more recently dictionary learning [125].

Sparsity in the spectral domain corresponds to the rank of a matrix. Analogous to ℓ_1 norm, nuclear norm (a.k.a. trace norm) defined as the sum of singular values is a convex relaxation of the rank function. Notably, nuclear norm minimization methods are shown effective in completing a partially observed low-rank matrix, namely matrix completion[24] and in recovering a low-rank matrix from sparse corruptions as in RPCA[27]). The key assumptions typically include uniform random support of observations/corruptions and that the underlying subspace needs to be *incoherent* (or close to orthogonal) against standard basis[32][114].

Motivating applications of matrix completion are recommendation systems (also called collaborative filtering in some literature)[14, 126], imputing missing DNA data [60], sensor network localization[123], structure from motion (SfM)[68] and etc. Similarly, many problems can be modeled in the framework of RPCA, e.g. foreground detection[64], image alignment[112], photometric stereo[149] in computer vision.

Since real data are noisy, robust extensions of matrix completion and RPCA have been proposed and rigorously analyzed[22, 155]. Their empirical performance however is not satisfactory in many of the motivating applications[119]. In particular, those with clear physical meanings on the matrix rank (e.g., SfM, sensor network localization and photometric stereo) should benefit from a hard constraint on rank and be solved better by matrix factorization[1]. This intuition essentially motivated our studies in Chapter 5, where we propose an algorithm to solve the difficult optimization with constraints on rank and ℓ_0 norm of sparse corruptions. In fact, matrix factorization has been success-

[1] where rank constraint is implicitly imposed by the inner dimension of matrix product.

fully adopted in a wide array of applications such as movie recommendation [87], SfM [68, 135] and NRSfM [111]. For a more complete list of matrix factorization's applications, we refer the readers to the reviews in [122] (for machine learning) and [46](for computer vision) and the references therein.

A fundamental limit of the matrix factorization approach is the lack of theoretical analysis. Notable exceptions include [84] which studies the unique recoverability from a combinatorial algebraic perspective and [76] that provides performance and convergence guarantee for the popular alternating least squares (ALS) method that solves matrix factorization. These two studies however do not generalize to noisy data. Our results in Chapter 2 (first appeared in [142]) are the first robustness analysis of matrix factorization/low-rank subspace model hence in some sense justified its good performance in real life applications.

1.2 Union-of-Subspace Model and Subspace Clustering

Building upon the now-well-understood low-rank and sparse models, researchers have started to consider more complex structures in data. The union-of-subspace model appears naturally when low-dimensional data are generated from different sources. As a mixture model, or more precisely a generalized hybrid linear model, the first problem to consider is to cluster the data points according to their subspace membership, namely, "subspace clustering".

Thanks to the prevalence of low-rank subspace models in applications (as we surveyed in the previous section), subspace clustering has been attracting increasing attention from diverse fields of study. For instance, subspaces may correspond to motion trajectories of different moving rigid objects[53], different communities in a social graph[77] or packet hop-count within each subnet in a computer network[59].

Existing methods on this problem include EM-like iterative algorithms [18, 137], algebraic methods (e.g., GPCA [140]), factorization [43], spectral clustering [35] as well as the latest Sparse Subspace Clustering (SSC)[53, 56, 124] and Low-Rank Represen-

tation (LRR)[96, 98]. While a number of these algorithms have theoretical guarantee, SSC is the only polynomial time algorithm that is guaranteed to work on a condition weaker than independent subspace. Moreover, prior to the technique in Chapter 3 (first made available online in [143] in November 2012), there has been no provable guarantee for any subspace clustering algorithm to work robustly under noise and model uncertainties, even though the robust variation of SSC and LRR have been the state-of-the-art on the Hopkins155 benchmark dataset[136] for quite a while.

In addition to the robustness results in Chapter 3, Chapter 4 focuses on developing a new algorithm that combines the advantages of LRR and SSC. Its results reveal new insights into both LRR and SSC as well as some new findings on the graph connectivity problem [104].

1.3 Structure of the Thesis

The chapters in this thesis are organized as follows.

In **Chapter 2 Stability of Matrix Factorization for Collaborative Filtering**, we study the stability *vis a vis* adversarial noise of matrix factorization algorithm for noisy and known-rank matrix completion. The results include stability bounds in three different evaluation metrics. Moreover, we apply these bounds to the problem of collaborative filtering under manipulator attack, which leads to useful insights and guidelines for collaborative filtering/recommendation system design. Part of the results in this chapter appeared in [142].

In **Chapter 3 Robust Subspace Clustering via Lasso-SSC**, we considers the problem of subspace clustering under noise. Specifically, we study the behavior of Sparse Subspace Clustering (SSC) when either adversarial or random noise is added to the unlabelled input data points, which are assumed to follow the union-of-subspace model. We show that a modified version of SSC is *provably effective* in correctly identifying the underlying subspaces, even with noisy data. This extends theoretical guarantee of this algorithm to the practical setting and provides justification to the success of SSC in

a class of real applications. Part of the results in this chapter appeared in [143].

In **Chapter 4 When LRR meets SSC: the separation-connectivity tradeoff**, we consider a slightly different notion of robustness for the subspace clustering problem: the connectivity of the constructed affinity graph for each subspace block. Ideally, the corresponding affinity matrix should be block diagonal with each diagonal block fully connected. Previous works such consider only the block diagonal shape[1] but not the connectivity, hence could not rule out the potential over-segmentation of subspaces. By combining SSC with LRR into LRSSC, and analyzing its performance, we find that the tradeoff between the ℓ_1 and nuclear norm penalty essentially trades off between separation (block diagonal) and connection density (implying connectivity). Part of the results in this chapter is submitted to NIPS[145] and is currently under review.

In **Chapter 5 PARSuMi: Practical Matrix Completion and Corruption Recovery with Explicit Modeling**, we identify and address the various weakness of nuclear norm-based approaches on real data by designing a practically working robust matrix completion algorithm. Specifically, we develop a Proximal Alternating Robust Subspace Minimization (PARSuMi) method to simultaneously handle missing data, sparse corruptions and dense noise. The alternating scheme explicitly exploits the rank constraint on the completed matrix and uses the ℓ_0 pseudo-norm directly in the corruption recovery step. While the method only converges to a stationary point, we demonstrate that its explicit modeling helps PARSuMi to work much more satisfactorily than nuclear norm based methods on synthetic and real data. In addition, this chapter also includes a comprehensive evaluation of existing methods for matrix factorization as well as their comparisons to the nuclear norm minimization-based convex methods, which is interesting on its own right. Part of the materials in this chapter is included in our manuscript [144] which is currently under review.

Finally, in **Chapter 6 Conclusion and Future Work**, we wrap up the thesis with a concluding discussions and then list the some open questions and potential future developments related to this thesis.

[1] also known as, self-expressiveness in [56] and subspace detection property in [124].

Chapter 2

Stability of Matrix Factorization for Collaborative Filtering

In this chapter, we study the stability *vis a vis* adversarial noise of matrix factorization algorithm for matrix completion. In particular, our results include: (I) we bound the gap between the solution matrix of the factorization method and the ground truth in terms of root mean square error; (II) we treat the matrix factorization as a subspace fitting problem and analyze the difference between the solution subspace and the ground truth; (III) we analyze the prediction error of individual users based on the subspace stability. We apply these results to the problem of collaborative filtering under manipulator attack, which leads to useful insights and guidelines for collaborative filtering system design. Part of the results in this chapter appeared in [142].

2.1 Introduction

Collaborative prediction of user preferences has attracted fast growing attention in the machine learning community, best demonstrated by the million-dollar Netflix Challenge. Among various models proposed, matrix factorization is arguably the most widely applied method, due to its high accuracy, scalability [132] and flexibility to incorporating domain knowledge [87]. Hence, not surprisingly, matrix factorization

is the centerpiece of most state-of-the-art collaborative filtering systems, including the winner of Netflix Prize [12]. Indeed, matrix factorization has been widely applied to tasks other than collaborative filtering, including structure from motion, localization in wireless sensor network, DNA microarray estimation and beyond. Matrix factorization is also considered as a fundamental building block of many popular algorithms in regression, factor analysis, dimension reduction, and clustering [122].

Despite the popularity of factorization methods, not much has been done on the theoretical front. In this chapter, we fill the blank by analyzing the stability *vis a vis* adversarial noise of the matrix factorization methods, in hope of providing useful insights and guidelines for practitioners to design and diagnose their algorithm efficiently.

Our main contributions are three-fold: In Section 2.3 we bound the gap between the solution matrix of the factorization method and the ground truth in terms of root mean square error. In Section 2.4, we treat the matrix factorization as a subspace fitting problem and analyze the difference between the solution subspace and the ground truth. This facilitates an analysis of the prediction error of individual users, which we present in Section 2.5. To validate these results, we apply them to the problem of collaborative filtering under manipulator attack in Section 2.6. Interestingly, we find that matrix factorization are robust to the so-called "targeted attack", but not so to the so-called "mass attack" unless the number of manipulators are small. These results agree with the simulation observations.

We briefly discuss relevant literatures. Azar et al. [4] analyzed asymptotic performance of matrix factorization methods, yet under stringent assumptions on the fraction of observation and on the singular values. Drineas et al. [51] relaxed these assumptions but it requires a few fully rated users – a situation that rarely happens in practice. Srebro [126] considered the problem of the generalization error of learning a low-rank matrix. Their technique is similar to the proof of our first result, yet applied to a different context. Specifically, they are mainly interested in *binary* prediction (i.e., "like/dislike") rather than recovering the *real-valued* ground-truth matrix (and its column subspace). In addition, they did not investigate the stability of the algorithm under noise and ma-

nipulators.

Recently, some alternative algorithms, notably StableMC [22] based on nuclear norm optimization, and OptSpace [83] based on gradient descent over the Grassmannian, have been shown to be stable *vis a vis* noise [22, 82]. However, these two methods are less effective in practice. As documented in Mitra et al. [101], Wen [146] and many others, matrix factorization methods typically outperform these two methods. Indeed, our theoretical results reassure these empirical observations, see Section 2.3 for a detailed comparison of the stability results of different algorithms.

2.2 Formulation

2.2.1 Matrix Factorization with Missing Data

Let the user ratings of items (such as movies) form a matrix Y, where each column corresponds to a user and each row corresponds to an item. Thus, the ij^{th} entry is the rating of item-i from user-j. The valid range of the rating is $[-k, +k]$. Y is assumed to be a rank-r matrix[1], so there exists a factorization of this rating matrix $Y = UV^T$, where $Y \in \mathbb{R}^{m\times n}$, $U \in \mathbb{R}^{m\times r}$, $V \in \mathbb{R}^{n\times r}$. Without loss of generality, we assume $m \leq n$ throughout the chapter.

Collaborative filtering is about to recover the rating matrix from a fraction of entries possibly corrupted by noise or error. That is, we observe $\widehat{Y}_{ij}$ for $(ij) \in \Omega$ the sampling set (assumed to be uniformly random), and $\widehat{Y} = Y + E$ being a corrupted copy of Y, and we want to recover Y. This naturally leads to the optimization program below:

$$\begin{aligned} \min_{U,V} \quad & \frac{1}{2}\left\| P_\Omega(UV^T - \widehat{Y})\right\|_F^2 \\ \text{subject to} \quad & \left|[UV^T]_{i,j}\right| \leq k, \end{aligned} \tag{2.1}$$

[1] In practice, this means the user's preference of movies are influenced by no more than r latent factors.

where P_Ω is the sampling operator defined to be:

$$[P_\Omega(Y)]_{i,j} = \begin{cases} Y_{i,j} & \text{if } (i,j) \in \Omega; \\ 0 & \text{otherwise.} \end{cases} \tag{2.2}$$

We denote the optimal solution $Y^* = U^* V^{*T}$ and the error $\Delta = Y^* - Y$.

2.2.2 Matrix Factorization as Subspace Fitting

As pointed out in Chen [37], an alternative interpretation of collaborative filtering is fitting the optimal r-dimensional subspace $\mathcal{N}$ to the sampled data. That is, one can reformulate (2.1) into an equivalent form[1]:

$$\min_N f(N) = \sum_i \|(I - \mathbb{P}_i) y_i\|^2 = \sum_i y_i^T (I - \mathbb{P}_i) y_i, \tag{2.3}$$

where y_i is the observed entries in the i^{th} column of Y, N is an $m \times r$ matrix representing an orthonormal basis[2] of $\mathcal{N}$, N_i is the restriction of N to the observed entries in column i, and $\mathbb{P}_i = N_i (N_i^T N_i)^{-1} N_i^T$ is the projection onto $\mathrm{span}(N_i)$.

After solving (2.3), we can estimate the full matrix in a column by column manner via (2.4). Here y_i^* denotes the full i^{th} column of recovered rank-r matrix Y^*.

$$y_i^* = N(N_i^T N_i)^{-1} N_i^T y_i = N \mathrm{pinv}(N_i) y_i. \tag{2.4}$$

Due to error term E, the ground truth subspace $\mathcal{N}^{gnd}$ can not be obtained. Instead, denote the optimal subspace of (2.1) (equivalently (2.3)) by $\mathcal{N}^*$, and we bound the gap between these two subspaces using Canonical angle. The canonical angle matrix Θ is an $r \times r$ diagonal matrix, with the i^{th} diagonal entry $\theta_i = \arccos \sigma_i((N^{gnd})^T N^*)$.

The error of subspace recovery is measured by $\rho = \|\sin \Theta\|_2$, justified by the fol-

[1] Strictly speaking, this is only equivalent to (2.1) without the box constraint. See the discussion in appendix for our justifications.

[2] It is easy to see $N = \mathrm{ortho}(U)$ for U in (2.1)

lowing properties adapted from Chapter 2 of Stewart and Sun [130]:

$$\begin{aligned} \|\mathbb{P}^{gnd} - \mathbb{P}^{\mathcal{N}^*}\|_F &= \sqrt{2}\|\sin\Theta\|_F, \\ \|\mathbb{P}^{gnd} - \mathbb{P}^{\mathcal{N}^*}\|_2 &= \|\sin\Theta\|_2 = \sin\theta_1. \end{aligned} \tag{2.5}$$

2.2.3 Algorithms

We focus on the stability of the *global optimal solution* of Problem (2.1). As Problem (2.1) is not convex, finding the global optimum is non-trivial in general. While this is certainly an important question, it is beyond the scope of this chapter. Instead, we briefly review some results on this aspect.

The simplest algorithm for (2.1) is perhaps the alternating least squares method (ALS) which alternatingly minimizes the objective function over U and V until convergence. More sophisticatedly, second-order algorithms such as Wiberg, Damped Newton and Levenberg Marquadt are proposed with better convergence rate, as surveyed in Okatani and Deguchi [109]. Specific variations for CF are investigated in Takács et al. [133] and Koren et al. [87]. Furthermore, Jain et al. [76] proposed a variation of the ALS method and show for the first time, factorization methods provably reach the global optimal solution under a similar condition as nuclear norm minimization based matrix completion[32].

From an empirical perspective, Mitra et al. [101] reported that the global optimum is often obtained in simulation and Chen [37] demonstrated satisfactory percentage of hits to global minimum from randomly initialized trials on a real data set. To add to the empirical evidence, we provide a comprehensive numerical evaluation of popular matrix factorization algorithms with noisy and ill-conditioned data matrices in Section 5.3 of Chapter 5. The results seem to imply that matrix factorization requires fundamentally smaller sample complexity than nuclear norm minimization-based approaches.

2.3 Stability

We show in this section that when *sufficiently many* entries are sampled, the global optimal solution of factorization methods is stable *vis a vis* noise – i.e., it recovers a matrix "close to" the ground-truth. This is measured by the root mean square error (RMSE):

$$\text{RMSE} = \frac{1}{\sqrt{mn}}\|Y^* - Y\| \tag{2.6}$$

Theorem 2.1. *There exists an absolute constant C, such that with probability at least $1 - 2\exp(-n)$,*

$$\text{RMSE} \leq \frac{1}{\sqrt{|\Omega|}}\|P_\Omega(E)\|_F + \frac{\|E\|_F}{\sqrt{mn}} + Ck\left(\frac{nr\log(n)}{|\Omega|}\right)^{\frac{1}{4}}.$$

Notice that when $|\Omega| \gg nr\log(n)$ the last term diminishes, and the RMSE is essentially bounded by the "average" magnitude of entries of E, i.e., the factorization methods are stable.

Comparison with related work

We recall similar RMSE bounds for StableMC of Candes and Plan [22] and OptSpace of Keshavan et al. [82]:

$$\textbf{StableMC: } \text{RMSE} \leq \sqrt{\frac{32\min(m,n)}{|\Omega|}}\|P_\Omega(E)\|_F + \frac{1}{\sqrt{mn}}\|P_\Omega(E)\|_F, \tag{2.7}$$

$$\textbf{OptSpace: } \text{RMSE} \leq C\kappa^2\frac{n\sqrt{r}}{|\Omega|}\|P_\Omega(E)\|_2. \tag{2.8}$$

Albeit the fact that these bounds are for different algorithms and under different assumptions (see Table 2.1 for details), it is still interesting to compare the results with Theorem 2.1. We observe that Theorem 2.1 is tighter than (2.7) by a scale of $\sqrt{\min(m,n)}$, and tighter than (2.8) by a scale of $\sqrt{n/\log(n)}$ in case of adversarial noise. However, the latter result is stronger when the noise is stochastic, due to the spectral norm used.

	Rank constraint	$Y_{i,j}$ constraint	σ constraint	incoherence	global optimal
Theorem 2.1	fixed rank	box constraint	no	no	assumed
OptSpace	fixed rank	regularization	condition number	weak	not necessary
NoisyMC	relaxed to trace	implicit	no	strong	yes

Table 2.1: Comparison of assumptions between stability results in our Theorem 2.1, OptSpace and NoisyMC

Compare with an Oracle

We next compare the bound with an oracle, introduced in Candes and Plan [22], that is assumed to know the ground-truth column space $\mathcal{N}$ *a priori* and recover the matrix by projecting the observation to $\mathcal{N}$ in the least squares sense column by column via (2.4). It is shown that RMSE of this oracle satisfies,

$$\text{RMSE} \approx \sqrt{1/|\Omega|}\|P_\Omega(E)\|_F. \tag{2.9}$$

Notice that Theorem 2.1 matches this oracle bound, and hence it is tight up to a constant factor.

2.3.1 Proof of Stability Theorem

We briefly explain the proof idea first. By definition, the algorithm finds the optimal rank-r matrices, measured in terms of the root mean square (RMS) on the *sampled* entries. To show this implies a small RMS on the *entire matrix*, we need to bound their gap

$$\tau(\Omega) \triangleq \Big|\frac{1}{\sqrt{|\Omega|}}\|P_\Omega(\widehat{Y} - Y^*)\|_F - \frac{1}{\sqrt{mn}}\|\widehat{Y} - Y^*\|_F\Big|.$$

To bound $\tau(\Omega)$, we require the following theorem.

Theorem 2.2. *Let* $\hat{\mathcal{L}}(X) = \frac{1}{\sqrt{|\Omega|}}\|P_\Omega(X - \widehat{Y})\|_F$ *and* $\mathcal{L}(X) = \frac{1}{\sqrt{mn}}\|X - \widehat{Y}\|_F$ *be the empirical and actual loss function respectively. Furthermore, assume entry-wise constraint* $\max_{i,j}|X_{i,j}| \le k$. *Then for all rank-*$r$ *matrices* X, *with probability greater*

than $1 - 2\exp(-n)$*, there exists a fixed constant* C *such that*

$$\sup_{X \in S_r} |\hat{\mathcal{L}}(X) - \mathcal{L}(X)| \leq Ck\Big(\frac{nr\log(n)}{|\Omega|}\Big)^{\frac{1}{4}}.$$

Indeed, Theorem 2.2 easily implies Theorem 2.1.

Proof of Theorem 2.1. The proof makes use of the fact that Y^* is the global optimal of (2.1).

$$\begin{aligned}
\text{RMSE} =& \frac{1}{\sqrt{mn}}\|Y^* - Y\|_F = \frac{1}{\sqrt{mn}}\|Y^* - \widehat{Y} + E\|_F \\
\leq& \frac{1}{\sqrt{mn}}|Y^* - \widehat{Y}\|_F + \frac{1}{\sqrt{mn}}\|E\|_F \\
\overset{(a)}{\leq}& \frac{1}{\sqrt{|\Omega|}}\|P_\Omega(Y^* - \widehat{Y})\|_F + \tau(\Omega) + \frac{1}{\sqrt{mn}}\|E\|_F \\
\overset{(b)}{\leq}& \frac{1}{\sqrt{|\Omega|}}\|P_\Omega(Y - \widehat{Y})\|_F + \tau(\Omega) + \frac{1}{\sqrt{mn}}\|E\|_F \\
=& \frac{1}{\sqrt{|\Omega|}}\|P_\Omega(E)\|_F + \tau(\Omega) + \frac{1}{\sqrt{mn}}\|E\|_F.
\end{aligned}$$

Here, (a) holds from definition of $\tau(\Omega)$, and (b) holds because Y^* is optimal solution of (2.1). Since $Y^* \in S_r$, applying Theorem 2.2 completes the proof. □

The proof of Theorem 2.2 is deferred to Appendix A.1 due to space constraints. The main idea, briefly speaking, is to bound, for a fixed $X \in S_r$,

$$\big|(\hat{\mathcal{L}}(X))^2 - (\mathcal{L}(X))^2\big| = \big|\frac{1}{|\Omega|}\|P_\Omega(X - \widehat{Y})\|_F^2 - \frac{1}{mn}\|X - \widehat{Y}\|_F^2\big|,$$

using Hoeffding's inequality for sampling without replacement; then bound $\big|\hat{\mathcal{L}}(X) - \mathcal{L}(X)\big|$ using

$$\big|\hat{\mathcal{L}}(X) - \mathcal{L}(X)\big| \leq \sqrt{\big|(\hat{\mathcal{L}}(X))^2 - (\mathcal{L}(X))^2\big|};$$

and finally, bound $\sup_{X \in S_r} |\hat{\mathcal{L}}(X) - \mathcal{L}(X)|$ using an $\epsilon-$net argument.

2.4 Subspace Stability

In this section we investigate the stability of recovered *subspace* using matrix factorization methods. Recall that matrix factorization methods assume that, in the idealized noiseless case, the preference of each user belongs to a low-rank subspace. Therefore, if this subspace can be readily recovered, then we can predict preferences of a new user without re-run the matrix factorization algorithms. We analyze the latter, prediction error on individual users, in Section 2.5.

To illustrate the difference between the stability of the recovered matrix and that of the recovered subspace, consider a concrete example in movie recommendation, where there are both honest users and malicious manipulators in the system. Suppose we obtain an output subspace N^* by (2.3) and the missing ratings are filled in by (2.4). If N^* is very "close" to ground truth subspace N, then all the predicted ratings for honest users will be good. On the other hand, the prediction error of the preference of the manipulators – who do not follow the low-rank assumption – can be large, which leads to a large error of the recovered matrix. Notice that we are only interested in predicting the preference of the honest users. Hence the subspace stability provides a more meaningful metric here.

2.4.1 Subspace Stability Theorem

Let $\mathcal{N}$,$\mathcal{M}$ and $\mathcal{N}^*$,$\mathcal{M}^*$ be the r-dimensional column space-row space pair of matrix Y and Y^* respectively. We'll denote the corresponding $m \times r$ and $n \times r$ orthonormal basis matrix of the vector spaces using N,M,N^*,M^*. Furthermore, Let Θ and Φ denote the canonical angles $\angle(\mathcal{N}^*, \mathcal{N})$ and $\angle(\mathcal{M}^*, \mathcal{M})$ respectively.

Theorem 2.3. *When Y is perturbed by additive error E and observed only on Ω, then there exists a Δ satisfying $\|\Delta\| \leq \sqrt{\frac{mn}{|\Omega|}}\|P_\Omega(E)\|_F + \|E\|_F + \sqrt{mn}\,|\tau(\Omega)|$, such that:*

$$\|\sin\Theta\| \leq \frac{\sqrt{2}}{\delta}\|(\mathbb{P}^{\mathcal{N}^\perp}\Delta)\|; \quad \|\sin\Phi\| \leq \frac{\sqrt{2}}{\delta}\|(\mathbb{P}^{\mathcal{M}^\perp}\Delta^T)\|,$$

where $\|\cdot\|$ is either the Frobenious norm or the spectral norm, and $\delta = \sigma_r^*$, i.e., the r^{th} largest singular value of the recovered matrix Y^*.

Furthermore, we can bound δ by:

$$\begin{cases} \sigma_r - \|\Delta\|_2 & \leq \delta \leq \quad \sigma_r + \|\Delta\|_2 \\ \sigma_r^{\tilde{Y}_{\mathcal{N}}} - \|\mathbb{P}^{\mathcal{N}^\perp}\Delta\|_2 & \leq \delta \leq \quad \sigma_r^{\tilde{Y}_{\mathcal{N}}} + \|\mathbb{P}^{\mathcal{N}^\perp}\Delta\|_2 \\ \sigma_r^{\tilde{Y}_{\mathcal{M}}} - \|\mathbb{P}^{\mathcal{M}^\perp}\Delta^T\|_2 & \leq \delta \leq \quad \sigma_r^{\tilde{Y}_{\mathcal{M}}} + \|\mathbb{P}^{\mathcal{M}^\perp}\Delta^T\|_2 \end{cases}$$

where $\tilde{Y}_{\mathcal{N}} = Y + \mathbb{P}^{\mathcal{N}}\Delta$ and $\tilde{Y}_{\mathcal{M}} = Y + (\mathbb{P}^{\mathcal{M}}\Delta^T)^T$.

Notice that in practice, as Y^* is the output of the algorithm, its r^{th} singular value δ is readily obtainable. Intuitively, Theorem 2.3 shows that the subspace sensitivity *vis a vis* noise depends on the singular value distribution of original matrix Y. A well-conditioned rank-r matrix Y can tolerate larger noise, as its r^{th} singular value is of the similar scale to $\|Y\|_2$, its largest singular value.

2.4.2 Proof of Subspace Stability

Proof of Theorem 2.3. In the proof, we use $\|\cdot\|$ when a result holds for both Frobenious norm and for spectral norm. We prove the two parts separately.

Part 1: Canonical Angles.

Let $\Delta = Y^* - Y$. By Theorem 2.1, we have $\|\Delta\| \leq \sqrt{\frac{mn}{|\Omega|}}\|P_\Omega(E)\|_F + \|E\|_F + \sqrt{mn}\,|\tau(\Omega)|$. The rest of the proof relates Δ with the deviation of spaces spanned by the top r singular vectors of Y and Y^* respectively. Our main tools are Weyl's Theorem and Wedin's Theorem (Lemma A.4 and A.5 in Appendix A.6).

We express singular value decomposition of Y and Y^* in block matrix form as in (A.10) and (A.11) of Appendix A.6, and set the dimension of Σ_1 and $\hat{\Sigma}_1$ to be $r \times r$. Recall, $\mathrm{rank}(Y) = r$, so $\Sigma_1 = \mathrm{diag}(\sigma_1, ..., \sigma_r)$, $\Sigma_2 = 0$, $\widehat{\Sigma}_1 = \mathrm{diag}(\sigma_1^{'}, ..., \sigma_r^{'})$. By setting $\widehat{\Sigma}_2$ to 0 we obtained Y', the nearest rank-r matrix to Y^*. Observe that $N^* = \hat{L}_1$, $M^* = (\hat{R}_1)^T$.

To apply Wedin's Theorem (Lemma A.5), we have the residual Z and S as follows:

$$Z = YM^* - N^*\widehat{\Sigma}_1, \qquad S = Y^T N^* - M^*\widehat{\Sigma}_1,$$

which leads to

$$\begin{aligned} \|Z\| &= \|(\widehat{Y} - \Delta)M^* - N^*\widehat{\Sigma}_1\| = \|\Delta M^*\|, \\ \|S\| &= \|(\widehat{Y} - \Delta)^T N^* - M^*\widehat{\Sigma}_1\| = \|\Delta^T N^*\|. \end{aligned}$$

Substitute this into the Wedin's inequality, we have

$$\sqrt{\|\sin\Phi\|^2 + \|\sin\Theta\|^2} \leq \frac{\sqrt{\|\Delta^T N'\|^2 + \|\Delta M'\|^2}}{\delta}, \tag{2.10}$$

where δ satisfies (A.12) and (A.13). Specifically, $\delta = \sigma_r^*$. Observe that Equation (2.10) implies

$$\|\sin\Theta\| \leq \frac{\sqrt{2}}{\delta}\|\Delta\|; \quad \|\sin\Phi\| \leq \frac{\sqrt{2}}{\delta}\|\Delta\|.$$

To reach the equations presented in the theorem, we can tighten the above bound by decomposing Δ into two orthogonal components.

$$Y^* = Y + \Delta = Y + \mathbb{P}^{\mathcal{N}}\Delta + \mathbb{P}^{\mathcal{N}^\perp}\Delta := \tilde{Y}^{\mathcal{N}} + \mathbb{P}^{\mathcal{N}^\perp}\Delta. \tag{2.11}$$

It is easy to see that column space of Y and $\tilde{Y}_{\mathcal{N}}$ are identical. So the canonical angle Θ between Y^* and Y are the same as that between Y^* and $\tilde{Y}_{\mathcal{N}}$. Therefore, we can replace Δ by $\mathbb{P}^{\mathcal{N}^\perp}\Delta$ to obtain the equation presented in the theorem. The corresponding result for row subspace follows similarly, by decomposing Δ^T to its projection on $\mathcal{M}$ and $\mathcal{M}^\perp$.

Part 2: Bounding δ.

We now bound δ, or equivalently σ_r^*. By Weyl's theorem (Lemma A.4),

$$|\delta - \sigma_r| < \|\Delta\|_2.$$

Moreover, Applying Weyl's theorem on Equation (2.11), we have

$$|\delta - \sigma_r^{\tilde{Y}_{\mathcal{N}}}| \leq \|\mathbb{P}_{\mathcal{N}^\perp}\Delta\|_2.$$

Similarly, we have

$$|\delta - \sigma_r^{\tilde{Y}_{\mathcal{M}}}| \leq \|\mathbb{P}_{\mathcal{M}^\perp}\Delta^T\|_2.$$

This establishes the theorem. □

2.5 Prediction Error of individual user

In this section, we analyze how confident we can predict the ratings of a new user $y \in \mathcal{N}^{gnd}$, based on the subspace recovered via matrix factorization methods. In particular, we bound the prediction $\|\tilde{y}^* - y\|$, where $\tilde{y}^*$ is the estimation from partial rating using (2.4), and y is the ground truth.

Without loss of generality, if the sampling rate is p, we assume observations occur in first pm entries, such that $y = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}$ with y_1 observed and y_2 unknown.

2.5.1 Prediction of y With Missing data

Theorem 2.4. *With all the notations and definitions above, and let N_1 denote the restriction of N on the observed entries of y. Then the prediction for $y \in \mathcal{N}^{gnd}$ has bounded performance:*

$$\|\tilde{y}^* - y\| \leq \left(1 + \frac{1}{\sigma_{min}}\right)\rho\|y\|,$$

where $\rho = \|\sin\Theta\|$ (see Theorem 2.3), σ_{min} is the smallest non-zero singular value of N_1 (r^{th} when N_1 is non-degenerate).

Proof. By (2.4), and recall that only the first pm entries are observed, we have

$$\tilde{y}^* = N \cdot \text{pinv}(N_1) y_1 := \begin{pmatrix} y_1 - \tilde{e}_1 \\ y_2 - \tilde{e}_2 \end{pmatrix} := y + \tilde{e}.$$

Let y^* be the vector obtained by projecting y onto subspace N, and denote $y^* = \begin{pmatrix} y_1^* \\ y_2^* \end{pmatrix} = \begin{pmatrix} y_1 - e_1 \\ y_2 - e_2 \end{pmatrix} = y - e$, we have:

$$\begin{aligned} \tilde{y}^* =& N \cdot \text{pinv}(N_1)(y_1^* + e_1) \\ =& N \cdot \text{pinv}(N_1) y_1^* + N \cdot \text{pinv}(N_1) e_1 = y^* + N \cdot \text{pinv}(N_1) e_1. \end{aligned}$$

Then

$$\begin{aligned} \|\tilde{y}^* - y\| =& \|y^* - y + N \cdot \text{pinv}(N_1) e_1\| \\ \leq& \|y^* - y\| + \frac{1}{\sigma_{min}} \|e_1\| \leq \rho \|y\| + \frac{1}{\sigma_{min}} \|e_1\|. \end{aligned}$$

Finally, we bound e_1 as follows

$$\|e_1\| \leq \|e\| = \|y - y^*\| \leq \|(\mathbb{P}^{gnd} - \mathbb{P}^{\mathcal{N}}) y\| \leq \rho \|y\|,$$

which completes the proof. □

Suppose $y \notin \mathcal{N}^{gnd}$ and $y = \mathbb{P}^{gnd} y + (I - \mathbb{P}^{gnd}) y := y^{gnd} + y^{gnd^\perp}$, then we have

$$\|e_1\| \leq \|(\mathbb{P}^{gnd} - \mathbb{P}^{\mathcal{N}}) y\| + \|y^{gnd^\perp}\| \leq \rho \|y\| + \|y^{gnd^\perp}\|,$$

which leads to

$$\|\tilde{y}^* - y^{gnd}\| \leq \left(1 + \frac{1}{\sigma_{min}}\right) \rho \|y\| + \frac{\|y^{gnd^\perp}\|}{\sigma_{min}}.$$

2.5.2 Bound on σ_{min}

To complete the above analysis, we now bound σ_{min}. Notice that in general σ_{min} can be arbitrarily close to zero, if N is "spiky". Hence we impose the strong incoherence property introduced in Candes and Tao [26] (see Appendix A.3 for the definition) to avoid such situation. Due to space constraint, we defer the proof of the following to the Appendix A.3.

Proposition 2.1. *If matrix Y satisfies strong incoherence property with parameter μ, then:*

$$\sigma_{min}(N_1) \geq 1 - \left(\frac{r}{m} + (1-p)\mu\sqrt{r}\right)^{\frac{1}{2}}.$$

For Gaussian Random Matrix

Stronger results on σ_{min} is possible for randomly generated matrices. As an example, we consider the case that $Y = UV$ where U, V are two Gaussian random matrices of size $m \times r$ and $r \times n$, and show that $\sigma_{min}(N_1) \approx \sqrt{p}$.

Proposition 2.2. *Let $G \in \mathbb{R}^{m\times r}$ have i.i.d. zero-mean Guassian random entries. Let N be its orthonormal basis*[1]*. Then there exists an absolute constant C such that with probability of at least $1 - Cn^{-10}$,*

$$\sigma_{min}(N_1) \geq \sqrt{\frac{k}{m}} - 2\sqrt{\frac{r}{m}} - C\sqrt{\frac{\log m}{m}}.$$

Due to space limit, the proof of Proposition 2.2 is deferred to the Appendix. The main idea is to apply established results about the singular values of Gaussian random matrix G [e.g., 45, 116, 121], then show that the orthogonal basis N of G is very close to G itself.

We remark that the bound on singular values we used has been generalized to random matrices following subgaussian [116] and log-concave distributions [95]. As such, the the above result can be easily generalized to a much larger class of random matrices.

[1] Hence N is also the orthonormal basis of any Y generated with G being its left multiplier.

2.6 Robustness against Manipulators

In this section, we apply our results to study the "profile injection" attacks on collaborative filtering. According to the empirical study of Mobasher et al. [102], matrix factorization, as a model-based CF algorithm, is more robust to such attacks compared to similarity-based CF algorithms such as kNN. However, as Cheng and Hurley [40] pointed out, it may not be a conclusive argument that model-based recommendation system is robust. Rather, it may due to the fact that that common attack schemes, effective to similarity based-approach, do not exploit the vulnerability of the model-based approach.

Our discovery is in tune with both Mobasher et al. [102] and Cheng and Hurley [40]. Specifically, we show that factorization methods are resilient to a class of common attack models, but are not so in general.

2.6.1 Attack Models

Depending on purpose, attackers may choose to inject "dummy profiles" in many ways. Models of different attack strategies are surveyed in Mobasher et al. [103]. For convenience, we propose to classify the models of attack into two distinctive categories: *Targeted Attack* and *Mass Attack*.

Targeted Attacks include average attack [89], segment attack and bandwagon attack [103]. The common characteristic of targeted attacks is that they *pretend* to be the honest users in all ratings except on a few targets of interest. Thus, each dummy user can be decomposed into:

$$e = e^{gnd} + s,$$

where $e^{gnd} \in \mathcal{N}$ and s is sparse.

Mass Attacks include random attack, love-hate attack [103] and others. The common characteristic of mass attacks is that they insert dummy users such that many en-

tries are manipulated. Hence, if we decompose a dummy user,

$$e = e^{gnd} + e^{gnd^\perp},$$

where $e^{gnd} = \mathbb{P}^{\mathcal{N}} e$ and $e^{gnd^\perp} = (I - \mathbb{P}^{\mathcal{N}})e \in \mathcal{N}^\perp$, then both components can have large magnitude. This is a more general model of attack.

2.6.2 Robustness Analysis

By definition, injected user profiles are column-wise: each dummy user corresponds to a corrupted column in the data matrix. For notational convenience, we re-arrange the order of columns into $[\, Y \mid E \,]$, where $Y \in \mathbb{R}^{m \times n}$ is of all honest users, and $E \in \mathbb{R}^{m \times n_e}$ contains all dummy users. As we only care about the prediction of honest users' ratings, we can, without loss of generality, set ground truth to be $[\, Y \mid E^{gnd} \,]$ and the additive error to be $[\, 0 \mid E^{gnd^\perp} \,]$. Thus, the recovery error $Z = [\, Y^* - Y \mid E^* - E^{gnd} \,]$.

Proposition 2.3. *Assume all conditions of Theorem 2.1 hold. Under "Targeted Attacks", there exists an absolute constant C, such that*

$$\text{RMSE} \leq 4k\sqrt{\frac{s_{max} n_e}{|\Omega|}} + Ck\left(\frac{(n+n_e) r \log(n+n_e)}{|\Omega|}\right)^{\frac{1}{4}}. \tag{2.12}$$

Here, s_{max} is maximal number of targeted items of each dummy user.

Proof. In the case of "Targeted Attacks", we have (recall that $k = \max_{(i,j)} |Y_{i,j}|$)

$$\|E^{gnd^\perp}\|_F < \sum_{i=1,\ldots,n_e} \|s_i\| \leq \sqrt{n_e s_{max} (2k)^2}.$$

Substituting this into Theorem 2.1 establishes the proposition. □

Remark 2.1. Proposition 2.3 essentially shows that matrix factorization approach is robust to the targeted attack model due to the fact that s_{max} is small. Indeed, if the sampling rate $|\Omega|/(m(n+n_e))$ is fixed, then RMSE converges to zero as m increases. This coincides with empirical results on Netflix data [12]. In contrast, similarity-based

algorithms (kNN) are extremely vulnerable to such attacks, due to the high similarity between dummy users and (some) honest users.

It is easy to see that the factorization method is less robust to mass attacks, simply because $\|E^{gnd^\perp}\|_F$ is not sparse, and hence s_{max} can be as large as m. Thus, the right hand side of (2.12) may not diminish. Nevertheless, as we show below, if the number of "Mass Attackers" does not exceed certain threshold, then the error will mainly concentrates on the E block. Hence, the prediction of the honest users is still acceptable.

Proposition 2.4. *Assume sufficiently random subspace N (i.e., Propostion 2.2 holds), above definition of "Mass Attacks", and condition number κ. If $n_e < \frac{\sqrt{n}}{\kappa^2 r}(\frac{\mathbf{E}|Y_{i,j}|^2}{k^2})$ and $|\Omega| = pm(n+n_e)$ satisfying $p > 1/m^{1/4}$, furthermore individual sample rate of each users is bounded within $[p/2, 3p/2]$,[1] then with probability of at least $1 - cm^{-10}$, the RMSE for honest users and for manipulators satisfies:*

$$\mathrm{RMSE}_Y \le C_1 \kappa k \left(\frac{r^3 \log(n)}{p^3 n}\right)^{1/4}, \qquad \mathrm{RMSE}_E \le \frac{C_2 k}{\sqrt{p}},$$

for some universal constant c, C_1 and C_2.

The proof of Proposition 2.4, deferred in the appendix, involves bounding the prediction error of each individual users with Theorem 2.4 and sum over Y block and E block separately. Subspace difference ρ is bounded with Theorem 2.1 and Theorem 2.3 together. Finally, σ_{min} is bounded via Proposition 2.2.

2.6.3 Simulation

To verify our robustness paradigm, we conducted simulation for both models of attacks. Y is generated by multiplying two 1000×10 gaussian random matrix and n_e attackers are appended to the back of Y. Targeted Attacks are produced by randomly choosing from a column of Y and assign 2 "push" and 2 "nuke" targets to 1 and -1 respectively. Mass Attacks are generated using uniform distribution. Factorization is performed using ALS. The results of the simulation are summarized in Figure 2.1 and 2.2. Figure 2.1

[1] This assumption is made to simplify the proof. It easily holds under i.i.d sampling.

compares the RMSE under two attack models. It shows that when the number of attackers increases, RMSE under targeted attack remains small, while RMSE under random attack significantly increases. Figure 2.2 compares RMSE_E and RMSE_Y under random attack. It shows that when n_e is small, $\mathrm{RMSE}_Y \ll \mathrm{RMSE}_E$. However, as n_e increases, RMSE_Y grows and eventually is comparable to RMSE_E. Both figures agree with our theoretic prediction. Additionally, from Figure 2.3, we can see a sharp transition in error level from honest user block on the left to the dummy user blocks on the right. This agrees with the prediction in Proposition 2.4 and the discussion in the beginning of Section 2.4. Lastly, Figure 2.4 illustrates the targeted attack version of Figure 2.2. From the curves, we can tell that while Proposition 2.3 bounds the total-$RMSE$, the gap between honest block and malicious block exists too. This leads to an even smaller manipulator impacts on honest users.

2.7 Chapter Summary

This chapter presented a comprehensive study of the stability of matrix factorization methods. The key results include a near-optimal stability bound, a subspace stability bound and a worst-case bound for individual columns. Then the theory is applied to the notorious manipulator problem in collaborative filtering, which leads to an interesting insight of MF's inherent robustness.

Matrix factorization is an important tool both for matrix completion task and for PCA with missing data. Yet, its practical success hinges on its stability – the ability to tolerate noise and corruption. The treatment in this chapter is a first attempt to understand the stability of matrix factorization, which we hope will help to guide the application of matrix factorization methods.

We list some possible directions to extend this research in future. In the theoretical front, the arguably most important open question is that under what conditions matrix factorization can reach a solution near global optimal. In the algorithmic front, we showed here that matrix factorization methods can be vulnerable to general manipu-

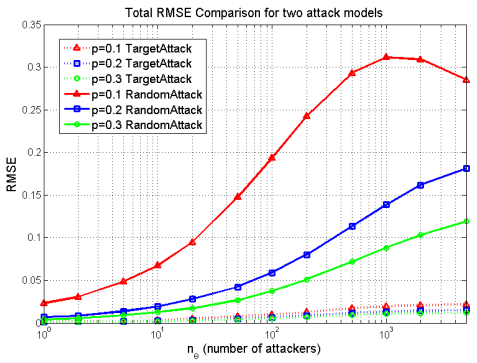


Figure 2.1: Comparison of two attack models.

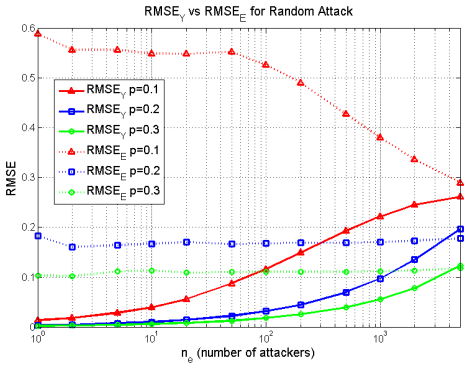


Figure 2.2: Comparison of RMSE_Y and RMSE_E under random attack.

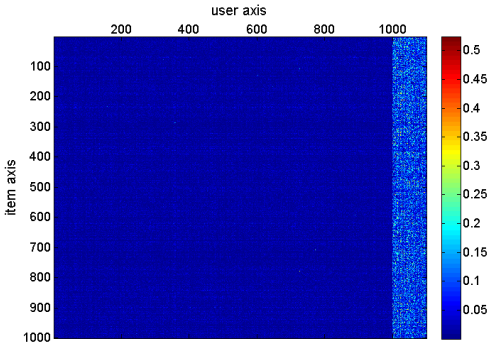


Figure 2.3: An illustration of error distribution for Random Attack, $n_e = 100$, $p = 0.3$.

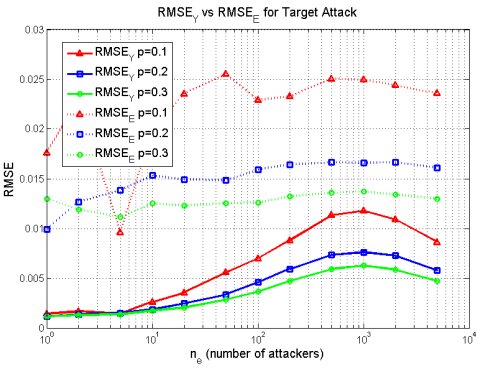


Figure 2.4: Comparison of $RMSE$ in Y-block and E-block for targeted attacks.

lators. Therefore, it is interesting to develop a robust variation of MF that provably handles arbitrary manipulators.

Later in Chapter D, we provide further study on matrix factorization, including empirical evaluation of existing algorithms, extensions to handle sparse corruptions and how the matrix factorization methods perform against nuclear norm minimization based approaches in both synthetic and real data.

Chapter 3

Robust Subspace Clustering via Lasso-SSC

This chapter considers the problem of subspace clustering under noise. Specifically, we study the behavior of Sparse Subspace Clustering (SSC) when either adversarial or random noise is added to the unlabelled input data points, which are assumed to lie in a union of low-dimensional subspaces. We show that a modified version of SSC is *provably effective* in correctly identifying the underlying subspaces, even with noisy data. This extends theoretical guarantee of this algorithm to the practical setting and provides justification to the success of SSC in a class of real applications. Part of the results in this chapter appeared in [143].

3.1 Introduction

Subspace clustering is a problem motivated by many real applications. It is now widely known that many high dimensional data including motion trajectories [42], face images [8], network hop counts [59], movie ratings [153] and social graphs [77] can be modelled as samples drawn from the *union* of multiple low-dimensional subspaces (illustrated in Figure 3.1). Subspace clustering, arguably the most crucial step to understand such data, refers to the task of clustering the data into their original subspaces

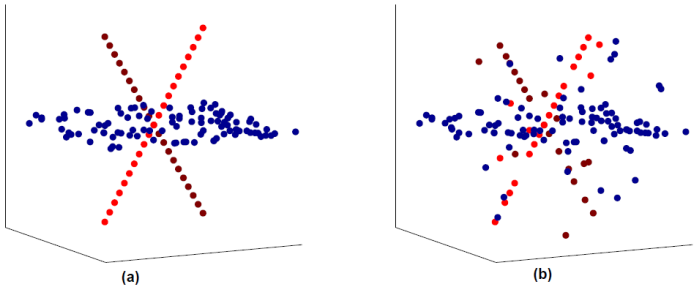


Figure 3.1: Exact (a) and noisy (b) data in union-of-subspace

and uncovers the underlying structure of the data. The partitions correspond to different rigid objects for motion trajectories, different people for face data, subnets for network data, like-minded users in movie database and latent communities for social graph.

Subspace clustering has drawn significant attention in the last decade and a great number of algorithms have been proposed, including K-plane [18], GPCA [140], Spectral Curvature Clustering [35], Low Rank Representation (LRR) [96] and Sparse Subspace Clustering (SSC) [53]. Among them, SSC is known to enjoy superb empirical performance, *even for noisy data*. For example, it is the state-of-the-art algorithm for motion segmentation on Hopkins155 benchmark [136]. For a comprehensive survey and comparisons, we refer the readers to the tutorial [139].

Effort has been made to explain the practical success of SSC. Elhamifar and Vidal [54] show that under certain conditions, *disjoint* subspaces (i.e., they are not overlapping) can be exactly recovered. Similar guarantee, under stronger "independent subspace" condition, was provided for LRR in a much earlier analysis[79]. The recent geometric analysis of SSC [124] broadens the scope of the results significantly to the case when subspaces can be overlapping. However, while these analyses advanced our understanding of SSC, one common drawback is that data points are assumed to be lying *exactly* in the subspace. This assumption can hardly be satisfied in practice. For example, motion trajectories data are only *approximately* rank-4 due to perspective distortion of camera.

In this chapter, we address this problem and provide the first theoretical analysis of

SSC with noisy or corrupted data. Our main result shows that a modified version of SSC (see (3.2)) when the magnitude of noise does not exceed a threshold determined by a geometric gap between *inradius* and *subspace incoherence* (see below for precise definitions). This complements the result of Soltanolkotabi and Candes [124] that shows the same geometric gap determines whether SSC succeeds for the noiseless case. Indeed, our results reduce to the noiseless results [124] when the noise magnitude diminishes.

While our analysis is based upon the geometric analysis in [124], the analysis is much more involved: In SSC, sample points are used as the dictionary for sparse recovery, and therefore noisy SSC requires analyzing noisy dictionary. This is a hard problem and we are not aware of any previous study that proposed guarantee in the case of noisy dictionary except Loh and Wainwright [100] in the high-dimensional regression problem. We also remark that our results on noisy SSC are *exact*, i.e., as long as the noise magnitude is smaller than the threshold, the obtained subspace recovery is *correct*. This is in sharp contrast to the majority of previous work on structure recovery for noisy data where stability/perturbation bounds are given – i.e., the obtained solution is *approximately* correct, and the approximation gap goes to zero only when the noise diminishes.

3.2 Problem Setup

Notations: We denote the uncorrupted data matrix by $Y \in \mathbb{R}^{n \times N}$, where each column of Y (normalized to unit vector) belongs to a union of L subspaces

$$\mathcal{S}_1 \cup \mathcal{S}_2 \cup ... \cup \mathcal{S}_L.$$

Each subspace $\mathcal{S}_\ell$ is of dimension d_ℓ and contains N_ℓ data samples with $N_1 + N_2 + ... + N_L = N$. We observe the noisy data matrix $X = Y + Z$, where Z is some arbitrary noise matrix. Let $Y^{(\ell)} \in \mathbb{R}^{n \times N_\ell}$ denote the selection of columns in Y that belongs to $\mathcal{S}_\ell$, and let the corresponding columns in X and Z be denoted by $X^{(\ell)}$ and $Z^{(\ell)}$. Without loss of generality, let $X = [X^{(1)}, X^{(2)}, ..., X^{(L)}]$ be ordered. In

addition, we use subscript "$-i$" to represent a matrix that excludes column i, e.g., $X_{-i}^{(\ell)} = [x_1^{(\ell)}, ..., x_{i-1}^{(\ell)}, x_{i+1}^{(\ell)}, ..., x_{N_\ell}^{(\ell)}]$. Calligraphic letters such as $\mathcal{X}, \mathcal{Y}_\ell$ represent the set containing all columns of the corresponding matrix (e.g., X and $Y^{(\ell)}$).

For any matrix X, $\mathcal{P}(X)$ represents the symmetrized convex hull of its columns, i.e., $\mathcal{P}(X) = \mathrm{conv}(\pm\mathcal{X})$. Also let $\mathcal{P}_{-i}^{(\ell)} := \mathcal{P}(X_{-i}^{(\ell)})$ and $\mathcal{Q}_{-i}^{(\ell)} := \mathcal{P}(Y_{-i}^{(\ell)})$ for short. $\mathbb{P}_{\mathcal{S}}$ and $\mathrm{Proj}_{\mathcal{S}}$ denote respectively the projection matrix and projection operator (acting on a set) to subspace $\mathcal{S}$. Throughout the chapter, $\|\cdot\|$ represents 2-norm for vectors and operator norm for matrices; other norms will be explicitly specified (e.g., $\|\cdot\|_1, \|\cdot\|_\infty$).

Method: Original SSC solves the linear program

$$\min_{c_i} \|c_i\|_1 \quad s.t. \quad x_i = X_{-i}c_i \tag{3.1}$$

for each data point x_i. Solutions are arranged into matrix $C = [c_1, ..., c_N]$, then spectral clustering techniques such as Ng et al. [106] are applied on the affinity matrix $W = |C| + |C|^T$. Note that when $Z \neq 0$, this method breaks down: indeed (3.1) may even be infeasible.

To handle noisy X, a natural extension is to relax the equality constraint in (3.1) and solve the following unconstrained minimization problem instead [56]:

$$\min_{c_i} \|c_i\|_1 + \frac{\lambda}{2}\|x_i - X_{-i}c_i\|^2. \tag{3.2}$$

We will focus on Formulation (3.2) in this chapter. Notice that (3.2) coincide with standard LASSO. Yet, since our task is subspace clustering, the analysis of LASSO (mainly for the task of support recovery) does not extend to SSC. In particular, existing literature for LASSO to succeed requires the dictionary X_{-i} to satisfy RIP [20] or the Null-space property [49], but neither of them is satisfied in the subspace clustering setup.[1]

In the subspace clustering task, there is no single "ground-truth" C to compare the

[1]There may exist two identical columns in X_{-i}, hence violate RIP for 2-sparse signal and has maximum incoherence $\mu(X_{-i}) = 1$.

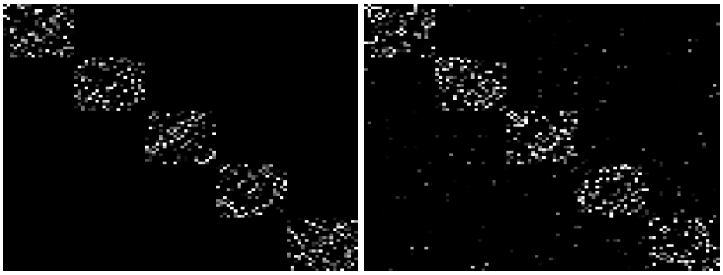


Figure 3.2: Illustration of LASSO-Subspace Detection Property/Self-Expressiveness Property. **Left:** SEP holds. **Right:** SEP is violated even though spectral clustering is likely to cluster this affinity graph perfectly into 5 blocks.

solution against. Instead, the algorithm succeeds if each sample is expressed as a linear combination of samples belonging to the same subspace, as the following definition states.

Definition 3.1 (LASSO Subspace Detection Property)**.** *We say subspaces* $\{\mathcal{S}_\ell\}_{\ell=1}^k$ *and noisy sample points* X *from these subspaces obey LASSO subspace detection property with* λ*, if and only if it holds that for all* i*, the optimal solution* c_i *to* (3.2) *with parameter* λ *satisfies:*

(1) c_i *is not a zero vector, (2) Nonzero entries of* c_i *correspond to only columns of* X *sampled from the same subspace as* x_i*.*

This property ensures that output matrix C and (naturally) affinity matrix W are exactly block diagonal with each subspace cluster represented by a disjoint block.[1] The property is illustrated in Figure 3.2. For convenience, we will refer to the second requirement alone as "*Self-Expressiveness Property*" (SEP), as defined in Elhamifar and Vidal [56].

Models of analysis: Our objective here is to provide sufficient conditions upon which the LASSO subspace detection properties hold in the following four models. Precise definition of the noise models will be given in Section 3.3.

[1]Note that this is a very strong condition. In general, spectral clustering does not require the exact block diagonal structure for perfect classifications (check Figure 3.6 in our simulation section for details).

- fully deterministic model
- deterministic data+random noise
- semi-random data+random noise
- fully random model.

3.3 Main Results

3.3.1 Deterministic Model

We start by defining two concepts adapted from Soltanolkotabi and Candes's original proposal.

Definition 3.2 (Projected Dual Direction[1])**.** *Let ν be the optimal solution to*

$$\max_{\nu} \langle x, \nu \rangle - \frac{1}{2\lambda}\nu^T\nu, \quad \textit{subject to:} \quad \|X^T\nu\|_\infty \leq 1;$$

and $\mathcal{S}$ is a low-dimensional subspace. The projected dual direction *v is defined as*

$$v(x, X, \mathcal{S}, \lambda) \triangleq \frac{\mathbb{P}_{\mathcal{S}}\nu}{\|\mathbb{P}_{\mathcal{S}}\nu\|}.$$

Definition 3.3 (Projected Subspace Incoherence Property)**.** *Compactly denote projected dual direction $v_i^{(\ell)} = v(x_i^{(\ell)}, X_{-i}^{(\ell)}, \mathcal{S}_\ell, \lambda)$ and $V^{(\ell)} = [v_1^{(\ell)}, ..., v_{N_\ell}^{(\ell)}]$. We say that vector set $\mathcal{X}_\ell$ is μ-incoherent to other points if*

$$\mu \geq \mu(\mathcal{X}_\ell) := \max_{y \in \mathcal{Y}\setminus\mathcal{Y}_\ell} \|{V^{(\ell)}}^T y\|_\infty.$$

Here, μ measures the incoherence between corrupted subspace samples $\mathcal{X}_\ell$ and clean data points in other subspaces. As $\|y\| = 1$ by definition, the range of μ is $[0, 1]$. In case of random subspaces in high dimension, μ is close to zero. Moreover, as we will see later, for deterministic subspaces and random data points, μ is proportional to their expected angular distance (measured by cosine of canonical angles).

[1]This definition relate to (3.8), the dual problem of (3.2), which we will define in the proof.

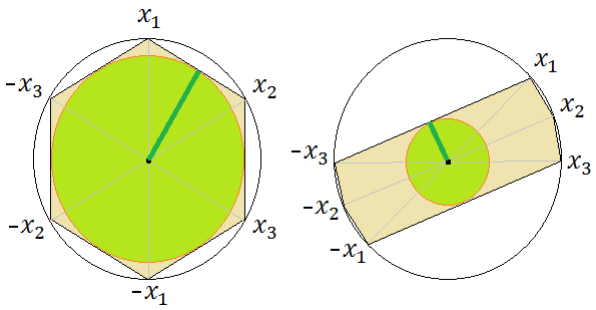


Figure 3.3: Illustration of inradius and data distribution.

Definition 3.2 and 3.3 are different from their original versions proposed in Soltanolkotabi and Candes [124] in that we require a projection to a particular subspace to cater to the analysis of the noise case.

Definition 3.4 (inradius)**.** *The inradius of a convex body $\mathcal{P}$, denoted by $r(\mathcal{P})$, is defined as the radius of the largest Euclidean ball inscribed in $\mathcal{P}$.*

The inradius of a $\mathcal{Q}_{-i}^{(\ell)}$ describes the distribution of the data points. Well-dispersed data lead to larger inradius and skewed/concentrated distribution of data have small inradius. An illustration is given in Figure 3.3.

Definition 3.5 (Deterministic noise model)**.** *Consider arbitrary additive noise Z to Y, each column z_i is characterized by the three quantities below:*

$$\delta := \max_i \|z_i\| \qquad \delta_1 := \max_{i,\ell} \|\mathbb{P}_{\mathcal{S}_\ell} z_i\| \qquad \delta_2 := \max_{i,\ell} \|\mathbb{P}_{\mathcal{S}_\ell^\perp} z_i\|$$

Theorem 3.1. *Under deterministic noise model, compactly denote*

$$\mu_\ell = \mu(\mathcal{X}_\ell), \qquad r_\ell := \min_{\{i:x_i \in \mathcal{X}_\ell\}} r(\mathcal{Q}_{-i}^{(\ell)}), \qquad r = \min_{\ell=1,\dots,L} r_\ell.$$

If $\mu_\ell < r_\ell$ for each $\ell = 1, ..., L$, furthermore

$$\delta \leq \min_{\ell=1,\dots,L} \frac{r(r_\ell - \mu_\ell)}{3r_\ell^2 + 8r_\ell + 2}$$

then LASSO subspace detection property holds for all weighting parameter λ in the

range

$$\begin{cases} \lambda > \dfrac{1}{(r-\delta_1)(1-3\delta)-3\delta-2\delta^2}, \\ \lambda < \dfrac{2r}{\delta^2(r+1)} \vee \min\limits_{\ell=1,..,L} \dfrac{r_\ell - \mu_\ell - 2\delta_1}{\delta(1+\delta)(3+2r_\ell-2\delta_1)}, \end{cases}$$

which is guaranteed to be non-empty.

Remark 3.1 (Noiseless case)**.** *When $\delta = 0$, i.e., there is no noise, the condition reduces to $\mu_\ell < r_\ell$, precisely the form in Soltanolkotabi and Candes [124]. However, the latter only works for the the exact LP formulation* (3.1)*, our result works for the (more robust) unconstrained LASSO formulation* (3.2) *for any $\lambda > \frac{1}{r}$.*

Remark 3.2 (Signal-to-Noise Ratio)**.** *Condition $\delta \leq \frac{r(r-\mu)}{3r^2+8r+2}$ can be interpreted as the breaking point under increasing magnitude of attack. This suggests that SSC by* (3.2) *is provably robust to arbitrary noise having signal-to-noise ratio (SNR) greater than $\Theta\big(\frac{1}{r(r-\mu)}\big)$. (Notice that $0 < r < 1$, we have $3r^2 + 8r + 2 = \Theta(1)$.)*

Remark 3.3 (Geometric Interpretation)**.** *The geometric interpretation of our results is give in Figure 3.4. On the left, Theorem 2.5 of Soltanolkotabi and Candes [124] suggests that the projection of external data points must fall inside the solid blue polygon, which is the intersection of halfspaces defined by dual directions (blue dots) that are tangent planes of the red inscribing sphere. On the right, the guarantee of Theorem 3.1 means that the whole red sphere (analogous to uncertainty set in Robust Optimization [13, 15]) of each external data point must fall inside the dashed red polygon, which is smaller than the blue polygon by a factor related to the noise level.*

Remark 3.4 (Matrix version of the algorithm)**.** *The theorem suggests there's a single λ that works for all x_i, X_{-i} in* (3.2)*. This makes it possible to extend the results to the compact matrix algorithm below*

$$\begin{aligned} \min_{C} \ & \|C\|_1 + \frac{\lambda}{2}\|X - XC_i\|_F^2 \\ \textit{s.t.} \quad & \operatorname{diag}(C) = 0, \end{aligned} \tag{3.3}$$

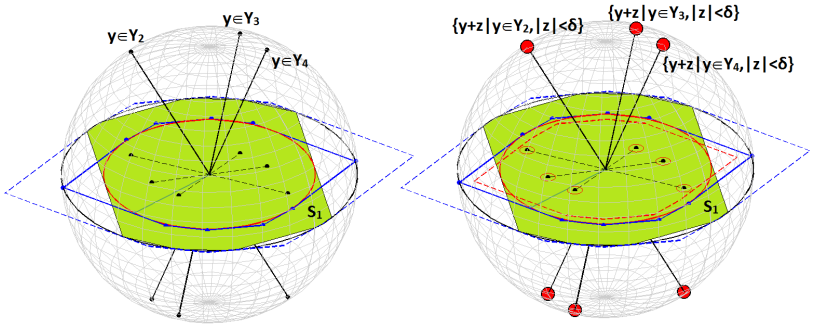


Figure 3.4: Geometric interpretation and comparison of the noiseless SSC (**Left**) and noisy Lasso-SSC (**Right**).

which can be solved numerically using alternating direction method of multipliers (ADMM) [17]. See the appendix for the details of the algorithm.

3.3.2 Randomized Models

We analyze three randomized models with increasing level of randomness.

- **Determinitic+Random Noise.** Subspaces and samples in subspace are fixed; noise is random (according to Definition 3.6).
- **Semi-random+Random Noise.** Subspace is deterministic, but samples in each subspace are drawn uniformly at random, noise is random.
- **Fully random.** Both subspace and samples are drawn uniformly at random; noise is also random.

Definition 3.6 (Random noise model)**.** *Our random noise model is defined to be any additive Z that is (1) columnwise iid; (2) spherical symmetric; and (3) $\|z_i\| \leq \delta$ with high probability.*

Example 3.1 (Gaussian noise)**.** *A good example of our random noise model is iid Gaussian noise. Let each entry $Z_{ij} \sim N(0, \sigma/\sqrt{n})$. It is known that*

$$\delta := \max_i \|z_i\| \leq \sqrt{1 + \frac{6 \log N}{n}} \sigma$$

with probability at least $1 - C/N^2$ *for some constant* C *(by Lemma B.5).*

Theorem 3.2 (Deterministic+Random Noise)**.** *Under random noise model, compactly denote* r_ℓ*,* r *and* μ_ℓ *as in Theorem 3.1, furthermore let*

$$\epsilon := \sqrt{\frac{6\log N + 2\log \max_\ell d_\ell}{n - \max_\ell d_\ell}} \leq \frac{C\log(N)}{\sqrt{n}}.$$

If $r > 3\epsilon/(1-6\epsilon)$ *and* $\mu_\ell < r_\ell$ *for all* $\ell = 1, ..., k$*, furthermore*

$$\delta < \min_{\ell=1,...,L} \frac{r_\ell - \mu_\ell}{3r_\ell + 6},$$

then with probability at least $1 - 7/N$*, LASSO subspace detection property holds for all weighting parameter* λ *in the range*

$$\begin{cases} \lambda > \dfrac{1}{(r-\delta\epsilon)(1-3\delta) - 3\delta - 2\delta^2}, \\ \lambda < \dfrac{2r}{\delta^2(r+1)} \vee \min\limits_{\ell=1,...,L} \dfrac{r_\ell - \mu_\ell - 2\delta\epsilon}{\epsilon\delta(1+\delta)(3+2r_\ell - 2\delta\epsilon)}, \end{cases}$$

which is guaranteed to be non-empty.

Remark 3.5 (Margin of error)**.** *Compared to Theorem 3.1, Theorem 3.2 considers a more benign noise which leads to a much stronger result. Observe that in the random noise case, the magnitude of noise that SSC can tolerate is proportional to* $r_\ell - \mu_\ell$ *– the difference of inradius and incoherence – which is the fundamental geometric gap that appears in the noiseless guarantee of Soltanolkotabi and Candes [124]. We call this gap the* ***Margin of error****.*

We now analyze this margin of error. We start from the semi-random model, where the distance between two subspaces is measured as follows.

Definition 3.7. *The* affinity *between two subspaces is defined by:*

$$\mathrm{aff}(\mathcal{S}_k, \mathcal{S}_\ell) = \sqrt{\cos^2\theta_{k\ell}^{(1)} + ... + \cos^2\theta_{k\ell}^{(\min(d_k,d_\ell))}},$$

where $\theta_{k\ell}^{(i)}$ is the i^{th} canonical angle between the two subspaces. Let U_k and U_ℓ be a set of orthonormal bases of each subspace, then $\operatorname{aff}(\mathcal{S}_k, \mathcal{S}_\ell) = \|U_k^T U_\ell\|_F$.

When data points are randomly sampled from each subspace, the geometric entity $\mu(\mathcal{X}_\ell)$ can be expressed using this (more intuitive) subspace affinity, which leads to the following theorem.

Theorem 3.3 (Semi-random+random noise)**.** *Suppose $N_\ell = \kappa_\ell d_\ell + 1$ data points are randomly chosen on each $\mathcal{S}_\ell$, $1 \leq \ell \leq L$. Use ϵ as in Theorem 3.2 and let $c(\kappa)$ be a positive constant that takes value $1/\sqrt{8}$ when κ is greater than some numerical constant κ_o. If*

$$\max_{k:k\neq\ell} t \log\left[LN_\ell(N_k+1)\right] \frac{\operatorname{aff}(\mathcal{S}_k, \mathcal{S}_\ell)}{\sqrt{d_k}} > c(\kappa_\ell)\sqrt{\frac{\log \kappa_\ell}{2}} \tag{3.4}$$

and $c(\kappa_\ell)\sqrt{\log \kappa_\ell / 2d_\ell} > 3\epsilon/(1-6\epsilon)$ for each ℓ, furthermore

$$\delta < \frac{1}{9} \min_\ell \left\{ \frac{c(\kappa_\ell)\sqrt{\log \kappa_\ell}}{\sqrt{2d_\ell}} - \max_{k:k\neq\ell} t \log\left[LN_\ell(N_k+1)\right] \frac{\operatorname{aff}(\mathcal{S}_k, \mathcal{S}_\ell)}{\sqrt{d_k d_\ell}} \right\},$$

then LASSO subspace detection property holds for some λ[1] with probability at least $1 - \frac{7}{N} - \sum_{\ell=1}^{L} N_\ell \exp(-\sqrt{d_\ell(N_\ell - 1)}) - \frac{1}{L^2}\sum_{k\neq\ell} \frac{1}{N_\ell(N_k+1)} \exp(-t/4)$.

Remark 3.6 (Overlapping subspaces)**.** *Similar to Soltanolkotabi and Candes [124], SSC can handle overlapping subspaces with noisy samples, as subspace affinity can take small positive value while still keeping the margin of error positive.*

Theorem 3.4 (Fully random model)**.** *Suppose there are L subspaces each with dimension d, chosen independently and uniformly at random. For each subspace, there are $\kappa d + 1$ points chosen independently and uniformly at random. Furthermore, each measurements are corrupted by iid Gaussian noise $\sim N(0, \sigma/\sqrt{n})$. Then for some absolute constant C, the LASSO subspace detection property holds for some λ with probability*

[1] The λ here (and that in Theorem 3.4) has a fixed non-empty range as in Theorem 3.1 and 3.2, which we omit due to space constraints.

Application	Cluster rank
3D motion segmentation [42]	$rank = 4$
Face clustering (with shadow) [8]	$rank = 9$
Diffuse photometric face [154]	$rank = 3$
Network topology discovery [59]	$rank = 2$
Hand writing digits [70]	$rank = 12$
Social graph clustering [77]	$rank = 1$

Table 3.1: Rank of real subspace clustering problems

at least $1 - \frac{C}{N} - Ne^{-\sqrt{\kappa}d}$ *if*

$$d < \frac{c^2(\kappa)\log\kappa}{12\log N}n \qquad \textit{and} \qquad \sigma < \frac{1}{18}\left(c(\kappa)\sqrt{\frac{\log\kappa}{2d}} - \sqrt{\frac{6\log N}{n}}\right).$$

Remark 3.7 (Trade-off between d and the margin of error)**.** *Theorem 3.4 extends our results to the paradigm where the subspace dimension grows linearly with the ambient dimension. Interestingly, it shows that the margin of error scales* $\tilde{\Theta}(\sqrt{1/d})$*, implying a tradeoff between* d *and robustness to noise. Fortunately, most interesting applications indeed have very low subspace-rank, as summarized in Table 3.1.*

Remark 3.8 (Robustness in the many-cluster setting)**.** *Another interesting observation is that the margin of error scales logarithmically with respect to* L*, the number of clusters (in both* $\log\kappa$ *and* $\log N$ *since* $N = L(\kappa d + 1)$*). This suggests that SSC is robust even if there are many clusters, and* $Ld \gg n$*.*

Remark 3.9 (Range of valid λ in the random setting)**.** *Substitute the bound of inradius* r *and subspace incoherence* μ *of fully random setting into the* λ*'s range of Theorem 3.3, we have the the valid range of* λ *is*

$$\frac{C_1\sqrt{d}}{\sqrt{\log\kappa}} < \lambda < \frac{C_2 n}{\sigma\sqrt{d\log(dL)}}, \tag{3.5}$$

for some constant C_1*,* C_2*. This again illustrates that the robustness is sensitive to* d *but not* L*.*

3.4 Roadmap of the Proof

In this section, we lay out the roadmap of the proof for Theorem 3.1 to 3.4. Instead of analyzing (3.2) directly, we consider an equivalent constrained version by introducing slack variables:

$$\mathbf{P}_0: \quad \min_{c_i, e_i} \|c_i\|_1 + \frac{\lambda}{2}\|e_i\|^2 \quad s.t. \quad x_i^{(\ell)} = X_{-i}c_i + e_i. \tag{3.6}$$

The constraint can be rewritten as

$$y_i^{(\ell)} + z_i^{(\ell)} = (Y_{-i} + Z_{-i})c_i + e_i. \tag{3.7}$$

The dual program of (3.6) is:

$$\mathbf{D}_0: \quad \max_{\nu} \langle x_i, \nu \rangle - \frac{1}{2\lambda}\nu^T\nu \quad s.t. \quad \|(X_{-i})^T\nu\|_\infty \leq 1. \tag{3.8}$$

Recall that we want to establish the conditions on noise magnitude δ, structure of the data (μ and r in deterministic model and affinity in semi-random model) and ranges of valid λ such that by Definition 3.1, the solution c_i is *non-trivial* and has support indices inside the column set $X_{-i}^{(\ell)}$ (i.e., satisfies SEP).

We focus on the proof of Theorem 3.1 and 3.2 and briefly explain the randomized models. Indeed, Theorem 3.3 and 3.4 follow directly by plugging to Theorem 3.2 the bound of r and μ from Soltanolkotabi and Candes [124] (with some modifications). The proof of Theorem 3.1 and 3.2 constitutes three main steps: (1) proving SEP, (2) proving non-trivialness, and (3) showing existence of proper λ.

3.4.1 Self-Expressiveness Property

We prove SEP by duality. First we establish a set of conditions on the optimal dual variable of D_0 corresponding to all primal solutions satisfying SEP. Then we construct such a dual variable ν as a certificate of proof.

3.4.1.1 Optimality Condition

Define general convex optimization:

$$\min_{c,e} \|c\|_1 + \frac{\lambda}{2}\|e\|^2 \qquad s.t. \quad x = Ac + e. \tag{3.9}$$

We state Lemma 3.1, which extends Lemma 7.1 in Soltanolkotabi and Candes [124]. The proof is deferred to the appendix.

Lemma 3.1. *Consider a vector $y \in \mathbb{R}^d$ and a matrix $A \in \mathbb{R}^{d\times N}$. If there exists a triplet (c, e, ν) obeying $y = Ac + e$ and c has support $S \subseteq T$, furthermore the dual certificate vector ν satisfies*

$$A_s^T\nu = sgn(c_S), \quad \nu = \lambda e, \quad \|A_{T\cap S^c}^T\nu\|_\infty \leq 1, \quad \|A_{T^c}^T\nu\|_\infty < 1,$$

then any optimal solution (c^*, e^*) to (3.9) obeys $c^*_{T^c} = 0$.

The next step is to apply Lemma 3.1 with $x = x_i^{(\ell)}$ and $A = X_{-i}$ and then construct a triplet (c, e, ν) such that dual certificate ν satisfying all conditions and c satisfies SEP. Then we can conclude that all optimal solutions of (3.6) satisfy SEP.

3.4.1.2 Construction of Dual Certificate

To construct the dual certificate, we consider the following *fictitious* optimization problem that explicitly requires that all feasible solutions satisfy SEP[1] (note that one can not solve such problem in practice without knowing the subspace clusters).

$$\begin{aligned} \mathbf{P}_1: \quad & \min_{c_i^{(\ell)}, e_i} \|c_i\|_1 + \frac{\lambda}{2}\|e_i\|^2 \\ & s.t. \quad y_i^{(\ell)} + z_i^{(\ell)} = (Y_{-i}^{(\ell)} + Z_{-i}^{(\ell)})c_i^{(\ell)} + e_i. \end{aligned} \tag{3.10}$$

This problem is feasible. Moreover, it turns out that the dual solution of this fictitious problem v is a good candidate as our dual certificate. Observe that v automatically

[1] To be precise, it's the corresponding $c_i = [0, ..., 0, (c_i^{(\ell)})^T, 0, ..., 0]^T$ that satisfies SEP.

satisfies the first three conditions in Lemma 3.1 and we are left to show that for all data point $x \in \mathcal{X} \setminus \mathcal{X}^\ell$,

$$|\langle x, \nu \rangle| < 1. \tag{3.11}$$

Let ν_1 and ν_2 be the projection of ν to subspace $\mathcal{S}_\ell$ and its complement respectively. The strategy is to provide an upper bound of $|\langle x, \nu \rangle|$ then impose the inequality on the upper bound.

$$\begin{aligned} |\langle x, \nu \rangle| =& |\langle y + z, \nu \rangle| \leq |\langle y, \nu_1 \rangle| + |\langle y, \nu_2 \rangle| + |\langle z, \nu \rangle| \\ \leq& \mu(\mathcal{X}_\ell)\|\nu_1\| + \|y\|\|\nu_2\||\cos(\angle(y, \nu_2))| + \|z\|\|\nu\||\cos(\angle(z, \nu))|. \end{aligned} \tag{3.12}$$

To complete the proof, we need to bound $\|\nu_1\|$ and $\|\nu_2\|$ and the two cosine terms (for random noise model). The proof makes use of the geometric properties of symmetric convex polytope and optimality of solution. See the appendix for the details.

3.4.2 Non-trivialness and Existence of λ

The idea is that when λ is large enough, trivial solution $c^* = 0$, $e^* = x_i^{(\ell)}$ can never occur. This is formalized by setting

$$\mathrm{OptVal}(\mathbf{D}_0) = \langle x_i^{(\ell)}, \nu \rangle - \frac{1}{2\lambda}\|\nu\|^2 < \frac{\lambda}{2}\|x_i^{(\ell)}\|^2. \tag{3.13}$$

Notice that (3.13) essentially requires that $\lambda > A$ and (3.12) requires $\lambda < B$ for some A and B. Hence, existence of a valid λ requires $A < B$, which leads to the condition on the error magnitude $\delta < C$ and completes the proof. While conceptually straightforward, the details of the proof are involved and left in the appendix due to space constraints.

3.4.3 Randomization

Our randomized results consider two types of randomization: *random noise* and *random data*. Random noise model improves the deterministic guarantee by exploiting the fact

that the directions of the noise are random. By the well-known bound on the area of spherical cap (Lemma B.4), the cosine terms in (3.12) diminishes when the ambient dimension grows. Similar advantage also appears in the bound of $\|\nu_1\|$ and $\|\nu_2\|$ and the existence of λ.

Randomization of data provides probabilistic bounds of inradius r and incoherence μ. The lower bound of inradius r follows from a lemma in the study of isotropy constant of symmetric convex body [2]. The upper bound of $\mu(\mathcal{X}_{-i}^{(\ell)})$ requires more effort. It involves showing that projected dual directions $v_i^{(\ell)}$ (see Definition 3.2) distributes uniformly on the subspace projection of the unit n-sphere, then applying the spherical cap lemma for all pairs of $(v_i^{(\ell)}, y)$. We defer the full proof in the appendix.

3.5 Numerical Simulation

To demonstrate the practical implications of our robustness guarantee for LASSO-SSC, we conduct three numerical experiments to test the effects of noise magnitude δ, subspace rank d and number of subspace L. To make it invariant to parameter, we scan through an exponential grid of λ ranging from $\sqrt{n} \times 10^{-2}$ to $\sqrt{n} \times 10^{3}$. In all experiments, ambient dimension $n = 100$, relative sampling $\kappa = 5$, subspace and data are drawn uniformly at random from unit sphere and then corrupted by Gaussian noise $Z_{ij} \sim N(0, \sigma/\sqrt{n})$. We measure the success of the algorithm by the relative violation of Self-Expressiveness Property defined below.

$$\mathrm{RelViolation}\,(C, \mathcal{M}) = \frac{\sum_{(i,j)\notin\mathcal{M}} |C|_{i,j}}{\sum_{(i,j)\in\mathcal{M}} |C|_{i,j}}$$

where $\mathcal{M}$ is the ground truth mask containing all (i, j) such that $x_i, x_j \in \mathcal{X}^{(\ell)}$ for some ℓ. Note that $\mathrm{RelViolation}\,(C, \mathcal{M}) = 0$ implies that SEP is satisfied. We also check that there is no all-zero columns in C, and the solution is considered trivial otherwise.

The simulation results confirm our theoretical findings. In particular, Figure 3.5 shows that LASSO subspace detection property is possible for a very large range of λ and the dependence on noise magnitude is roughly $1/\sigma$ as remarked in (3.5). In addition,

the sharp contrast of Figure 3.8 and 3.7 demonstrates precisely our observations on the sensitivity of d and L in Remark 3.7 and 3.8.

A remark on numerical algorithms: For fast computation, we use ADMM implementation of LASSO solver[1]. It has complexity proportional to problem size and convergence guarantee [17]. We also implement a simple solver for the matrix version SSC (3.3) which is consistently faster than the column-by-column LASSO version. Details of the algorithm and its favorable empirical comparisons are given in the appendix.

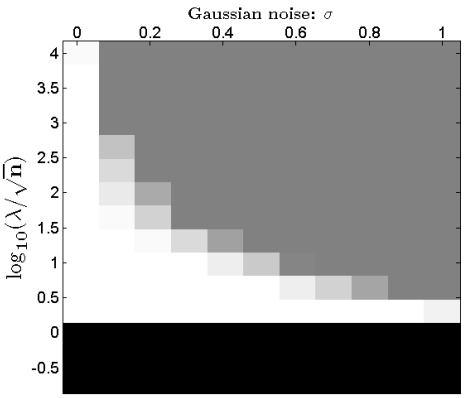


Figure 3.5: Exact recovery under noise. Simulated with $n = 100, d = 4, L = 3, \kappa = 5$ with increasing Gaussian noise $N(0, \sigma/\sqrt{n})$. **Black:** trivial solution ($C = 0$); **Gray:** RelViolation > 0.1; **White:** RelViolation $= 0$.

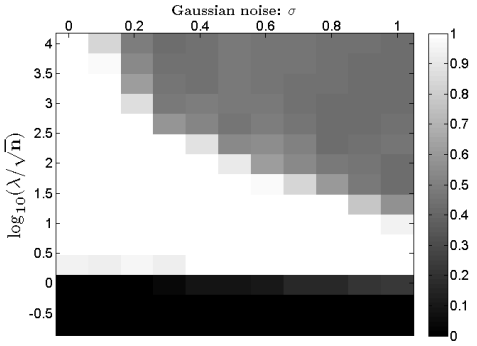


Figure 3.6: Spectral clustering accuracy for the experiment in Figure 3.5. The rate of accurate classification is represented in grayscale. White region means perfect classification. It is clear that exact subspace detection property (Definition 3.1) is not necessary for perfect classification.

3.6 Chapter Summary

We presented the first theoretical analysis for noisy subspace clustering problem that is of great practical interests. We showed that the popular SSC algorithm *exactly* (not approximately) succeeds even in the noisy case, which justified its empirical success on real problems. In addition, we discovered a fundamental trade-off between robustness to noise and the subspace dimension, and we found that robustness is insensitive

[1] Freely available at:
http://www.stanford.edu/~boyd/papers/admm/

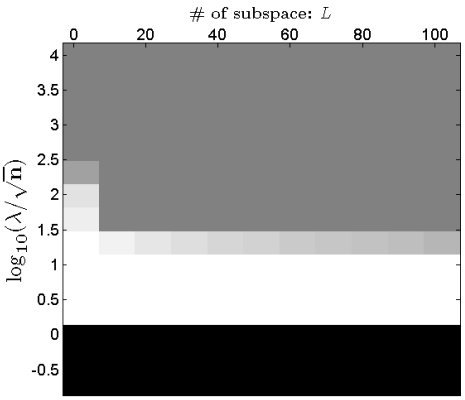


Figure 3.7: Effects of number of subspace L. Simulated with $n = 100, d = 2, \kappa = 5, \sigma = 0.2$ with increasing L. **Black:** trivial solution ($C = 0$); **Gray:** $\mathrm{RelViolation} > 0.1$; **White:** $\mathrm{RelViolation} = 0$. Note that even at the point when $dL = 200$(subspaces are highly dependent), subspace detection property holds for a large range of λ.

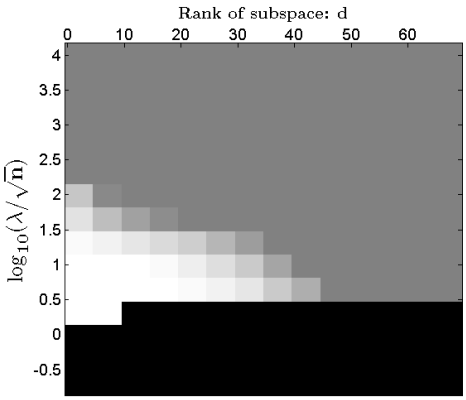


Figure 3.8: Effects of cluster rank d. Simulated with $n = 100, L = 3, \kappa = 5, \sigma = 0.2$ with increasing d. **Black:** trivial solution ($C = 0$); **Gray:** $\mathrm{RelViolation} > 0.1$; **White:** $\mathrm{RelViolation} = 0$. Observe that beyond a point, subspace detection property is not possible for any λ.

to the number of subspaces. Our analysis hence reveals fundamental relationships of robustness, number of samples and dimension of the subspace. These results lead to new theoretical understanding of SSC, as well as provides guidelines for practitioners and application-level researchers to judge whether SSC could possibly work well for their respective applications.

Open problems for subspace clustering include the graph connectivity problem raised by Nasihatkon and Hartley [104] (which we will talk about more in Chapter 4), missing data problem (a first attempt by Eriksson et al. [59], but requires an unrealistic number of data), sparse corruptions on data and others. One direction closely related to this chapter is to introduce a more practical metric of success. As we illustrated in this chapter, subspace detection property is not necessary for perfect clustering. In fact from a pragmatic point of view, even perfect clustering is not necessary. Typical applications allow for a small number of misclassifications. It would be interesting to see whether stronger robustness results can be obtained for a more practical metric of success.

Chapter 4

When LRR Meets SSC: the Separation-Connectivity Tradeoff

We continue to study the problem of *subspace clustering* in this chapter. The motivation deviates from the robustness to noise, but instead address the known weakness of SSC: the constructed graph may be too sparse within a single class.

This is the complete opposite of another successful algorithm termed Low-Rank Representation (LRR) that exploits the the same intuition of "Self-Expressiveness" as SSC. LRR often yields a very dense graph, as it minimizes nuclear norm (aka trace norm) to promote a low-rank structure in contract to SSC that minimizes the vector ℓ_1 norm of the representation matrix to induce sparsity.

We propose a new algorithm, termed Low-Rank Sparse Subspace Clustering (LRSSC), by combining SSC and LRR, and develops theoretical guarantees of when the algorithm succeeds. The results reveal interesting insights into the strength and weakness of SSC and LRR and demonstrate how LRSSC can take the advantages of both methods in preserving the "Self-Expressiveness Property" and "Graph Connectivity" at the same time. Part of the materials in this chapter is included in our submission[145].

4.1 Introduction

As discussed in the previous chapter, the wide array of problems that assume the structure of a union of low-rank subspaces has motivated various researchers to propose algorithms for the subspace clustering problem. Among these algorithms, Sparse Subspace Clustering (SSC) [53], Low Rank Representation (LRR) [98], based on minimizing the nuclear norm and ℓ_1 norm of the representation matrix respectively, remain the top performers on the Hopkins155 motion segmentation benchmark dataset[136]. Moreover, they are among the few subspace clustering algorithms supported with theoretic guarantees: Both algorithms are known to succeed when the subspaces are independent [98, 140]. Later, [56] showed that subspace being disjoint is sufficient for SSC to succeed[1], and [124] further relaxed this condition to include some cases of overlapping subspaces. Robustness of the two algorithms has been studied too. Liu et. al. [97] showed that a variant of LRR works even in the presence of some arbitrarily large outliers, while Wang and Xu [143] provided both deterministic and randomized guarantees for SSC when data are noisy or corrupted.

Despite LRR and SSC's success, there are questions unanswered. LRR has never been shown to succeed other than under the very restrictive "independent subspace" assumption. SSC's solution is sometimes too sparse that the affinity graph of data from a single subspace may not be a connected body [104]. Moreover, as our experiment with Hopkins155 data shows, the instances where SSC fails are often different from those that LRR fails. Hence, a natural question is whether combining the two algorithms lead to a better method, in particular since the underlying representation matrix we want to recover is *both low-rank and sparse* simultaneously.

In this chapter, we propose Low-Rank Sparse Subspace Clustering (LRSSC), which minimizes a weighted sum of nuclear norm and vector 1-norm of the representation matrix. We show theoretical guarantees for LRSSC that strengthen the results in [124]. The statement and proof also shed insight on why LRR requires independence assump-

[1] Disjoint subspaces only intersect at the origin. It is a less restrictive assumption comparing to independent subspaces, e.g., 3 coplanar lines passing the origin are not independent, but disjoint.

tion. Furthermore, the results imply that there is a fundamental trade-off between the interclass separation and the intra-class connectivity. Indeed, our experiment shows that LRSSC works well in cases where data distribution is skewed (graph connectivity becomes an issue for SSC) and subspaces are not independent (LRR gives poor separation). These insights would be useful when developing subspace clustering algorithms and applications. We remark that in the general regression setup, the simultaneous nuclear norm and 1-norm regularization has been studied before [115]. However, our focus is on the subspace clustering problem, and hence the results and analysis are completely different.

4.2 Problem Setup

Notations: We denote the data matrix by $X \in \mathbb{R}^{n\times N}$, where each column of X (normalized to unit vector) belongs to a union of L subspaces

$$\mathcal{S}_1 \cup \mathcal{S}_2 \cup ... \cup \mathcal{S}_L.$$

Each subspace ℓ contains N_ℓ data samples with $N_1 + N_2 + ... + N_L = N$. We observe the noisy data matrix X. Let $X^{(\ell)} \in \mathbb{R}^{n\times N_\ell}$ denote the selection (as a set and a matrix) of columns in X that belong to $\mathcal{S}_\ell \subset \mathbb{R}^n$, which is an d_ℓ-dimensional subspace. Without loss of generality, let $X = [X^{(1)}, X^{(2)}, ..., X^{(L)}]$ be ordered. In addition, we use $\|\cdot\|$ to represent Euclidean norm (for vectors) or spectral norm (for matrices) throughout the chapter.

Method: We solve the following convex optimization problem

$$\textbf{LRSSC}: \qquad \min_{C} \ \|C\|_* + \lambda\|C\|_1 \quad s.t. \quad X = XC, \quad \mathrm{diag}(C) = 0. \tag{4.1}$$

Spectral clustering techniques (e.g., [106]) are then applied on the affinity matrix $W = |C| + |C|^T$ where C is the solution to (4.1) to obtain the final clustering and $|\cdot|$ is the elementwise absolute value.

Criterion of success: In the subspace clustering task, as opposed to compressive sensing or matrix completion, there is no "ground-truth" C to compare the solution against. Instead, the algorithm succeeds if each sample is expressed as a linear combination of samples belonging to the *same subspace*, i.e., the output matrix C are *block diagonal* (up to appropriate permutation) with each subspace cluster represented by a disjoint block. Formally, we have the following definition.

Definition 4.1 (Self-Expressiveness Property (SEP))**.** *Given subspaces $\{\mathcal{S}_\ell\}_{\ell=1}^L$ and data points X from these subspaces, we say a matrix C obeys Self-Expressiveness Property, if the nonzero entries of each c_i (i^{th} column of C) corresponds to only those columns of X sampled from the same subspace as x_i.*

Note that the solution obeying SEP alone does not imply the clustering is correct, since each block may not be fully connected. This is the so-called "graph connectivity" problem studied in [104]. On the other hand, failure to achieve SEP does not necessarily imply clustering error either, as the spectral clustering step may give a (sometimes perfect) solution even when there are connections between blocks. Nevertheless, SEP is the condition that verifies the design intuition of SSC and LRR. Notice that if C obeys SEP and each block is connected, we immediately get the correct clustering.

4.3 Theoretic Guanratees

4.3.1 The Deterministic Setup

Before we state our theoretical results for the deterministic setup, we need to define a few quantities.

Definition 4.2 (Normalized dual matrix set)**.** *Let $\{\Lambda_1(X)\}$ be the set of optimal solutions to*

$$\max_{\Lambda_1,\Lambda_2,\Lambda_3} \langle X, \Lambda_1 \rangle \quad \textit{s.t.} \quad \|\Lambda_2\|_\infty \leq \lambda,\ \|X^T\Lambda_1 - \Lambda_2 - \Lambda_3\| \leq 1,\ \mathrm{diag}^\perp(\Lambda_3) = 0,$$

where $\|\cdot\|_\infty$ *is the vector* ℓ_∞ *norm and* $\mathrm{diag}^\perp$ *selects all the off-diagonal entries. Let* $\Lambda^* = [\nu_1^*, ..., \nu_N^*] \in \{\Lambda_1(X)\}$ *obey* $\nu_i^* \in \mathrm{span}(X)$ *for every* $i = 1, ..., N$.[1] *For every* $\Lambda = [\nu_1, ..., \nu_N] \in \{\Lambda_1(X)\}$, *we define* normalized dual matrix V *for* X *as*

$$V(X) \triangleq \left[\frac{\nu_1}{\|\nu_1^*\|}, ..., \frac{\nu_N}{\|\nu_N^*\|}\right],$$

and the normalized dual matrix set $\{V(X)\}$ *as the collection of* $V(X)$ *for all* $\Lambda \in \{\Lambda_1(X)\}$.

Definition 4.3 (Minimax subspace incoherence property)**.** *Compactly denote* $V^{(\ell)} = V(X^{(\ell)})$. *We say the vector set* $X^{(\ell)}$ *is* μ*-incoherent to other points if*

$$\mu \geq \mu(X^{(\ell)}) := \min_{V^{(\ell)} \in \{V^{(\ell)}\}} \max_{x \in X \setminus X^{(\ell)}} \|{V^{(\ell)}}^T x\|_\infty.$$

The incoherence μ in the above definition measures how *separable* the sample points in $\mathcal{S}_\ell$ are against sample points in other subspaces (small μ represents more separable data). Our definition differs from Soltanokotabi and Candes's definition of subspace incoherence [124] in that it is defined as a minimax over all possible dual directions. It is easy to see that μ-incoherence in [124, Definition 2.4] implies μ-minimax-incoherence as their dual direction are contained in $\{V(X)\}$. In fact, in several interesting cases, μ can be significantly smaller under the new definition. We illustrate the point with the two examples below and leave detailed discussions in the appendix.

Example 4.1 (Independent Subspace)**.** Suppose the subspaces are independent, i.e., $\dim(\mathcal{S}_1 \oplus ... \oplus \mathcal{S}_L) = \sum_{\ell=1,...,L} \dim(\mathcal{S}_\ell)$, then all $X^{(\ell)}$ are 0-incoherent under our Definition 4.3. This is because for each $X^{(\ell)}$ one can always find a dual matrix $V^{(\ell)} \in \{V^{(\ell)}\}$ whose column space is orthogonal to the span of all other subspaces. To contrast, the incoherence parameter according to Definition 2.4 in [124] will be a positive value, potentially large if the angles between subspaces are small.

[1] If this is not unique, pick the one with least Frobenious norm.

Example 4.2 (Random except 1 subspace)**.** Suppose we have L disjoint 1-dimensional subspaces in $\mathbb{R}^n$ ($L > n$). $\mathcal{S}_1, ..., \mathcal{S}_{L-1}$ subspaces are randomly drawn. $\mathcal{S}_L$ is chosen such that its angle to one of the $L-1$ subspace, say $\mathcal{S}_1$, is $\pi/6$. Then the incoherence parameter $\mu(X^{(L)})$ defined in [124] is at least $\cos(\pi/6)$. However under our new definition, it is not difficult to show that $\mu(X^{(L)}) \leq 2\sqrt{\frac{6\log(L)}{n}}$ with high probability[1].

The result also depends on the smallest singular value of a rank-d matrix (denoted by σ_d) and the inradius of a convex body as defined below.

Definition 4.4 (inradius)**.** *The inradius of a convex body $\mathcal{P}$, denoted by $r(\mathcal{P})$, is defined as the radius of the largest Euclidean ball inscribed in $\mathcal{P}$.*

The smallest singular value and inradius measure how *well-represented* each subspace is by its data samples. Small inradius/singular value implies either insufficient data, or skewed data distribution, in other word, it means that the subspace is "*poorly represented*". Now we may state our main result.

Theorem 4.1 (LRSSC)**.** *Self-expressiveness property holds for the solution of* (4.1) *on the data X if there exists a weighting parameter λ such that for all $\ell = 1, ..., L$, one of the following two conditions holds:*

$$\mu(X^{(\ell)})(1+\lambda\sqrt{N_\ell}) < \lambda \min_k \sigma_{d_\ell}(X^{(\ell)}_{-k}), \tag{4.2}$$

$$\text{or} \quad \mu(X^{(\ell)})(1+\lambda) < \lambda \min_k r(\text{conv}(\pm X^{(\ell)}_{-k})), \tag{4.3}$$

where X_{-k} denotes X with its k^{th} column removed and $\sigma_{d_\ell}(X^{(\ell)}_{-k})$ represents the d^{th}_ℓ (smallest non-zero) singular value of the matrix $X^{(\ell)}_{-k}$.

We briefly explain the intuition of the proof. The theorem is proven by duality. First we write out the dual problem of (4.1),

Dual LRSSC :

$$\max_{\Lambda_1,\Lambda_2,\Lambda_3} \langle X, \Lambda_1\rangle \quad \text{s.t. } \|\Lambda_2\|_\infty \leq \lambda,\ \|X^T\Lambda_1 - \Lambda_2 - \Lambda_3\| \leq 1,\ \text{diag}^\perp(\Lambda_3) = 0.$$

[1]The full proof is given in the Appendix. Also it is easy to generalize this example to d-dimensional subspaces and to "random except K subspaces".

This leads to a set of optimality conditions, and leaves us to show the existence of a dual certificate satisfying these conditions. We then construct two levels of fictitious optimizations (which is the main novelty of the proof) and *construct a dual certificate from the dual solution of the fictitious optimization problems*. Under condition (4.2) and (4.3), we establish this dual certifacte meets all optimality conditions, hence certifying that SEP holds. Due to space constraints, we defer the detailed proof to the appendix and focus on the discussions of the results in the main text.

Remark 4.1 (SSC)**.** Theorem 4.1 can be considered a generalization of Theorem 2.5 of [124]. Indeed, when $\lambda \to \infty$, (4.3) reduces to the following

$$\mu(X^{(\ell)}) < \min_k r(\text{conv}(\pm X^{(\ell)}_{-k})).$$

The readers may observe that this is exactly the same as Theorem 2.5 of [124], with the only difference being the definition of μ. Since our definition of $\mu(X^{(\ell)})$ is tighter (i.e., smaller) than that in [124], our guarantee for SSC is indeed stronger. Theorem 4.1 also implies that the good properties of SSC (such as overlapping subspaces, large dimension) shown in [124] are also valid for LRSSC for a range of λ greater than a threshold.

To further illustrate the key difference from [124], we describe the following scenario.

Example 4.3 (Correlated/Poorly Represented Subspaces)**.** Suppose the subspaces are poorly represented, i.e., the inradius r is small. If furthermore, the subspaces are highly correlated, i.e., canonical angles between subspaces are small, then the subspace incoherence μ' defined in [124] can be quite large (close to 1). Thus, the succeed condition $\mu' < r$ presented in [124] is violated. This is an important scenario because real data such as those in Hopkins155 and Extended YaleB often suffer from both problems, as illustrated in [57, Figure 9 & 10]. Using our new definition of incoherence μ, as long as the subspaces are "sufficiently independent"[1] (regardless of their correlation) μ will

[1]Due to space constraint, the concept is formalized in appendix.

assume very small values (e.g., Example 4.2), making SEP possible even if r is small, namely when subspaces are poorly represented.

Remark 4.2 (LRR)**.** The guarantee is the strongest when $\lambda \to \infty$ and becomes superficial when $\lambda \to 0$ unless subspaces are independent (see Example 4.1). This seems to imply that the "independent subspace" assumption used in [97, 98] to establish sufficient conditions for LRR (and variants) to work is unavoidable.[1] On the other hand, for each problem instance, there is a λ^* such that whenever $\lambda > \lambda^*$, the result satisfies SEP, so we should expect phase transition phenomenon when tuning λ.

Remark 4.3 (A tractable condition)**.** Condition (4.2) is based on singular values, hence is computationally tractable. In contrast, the verification of (4.3) or the deterministic condition in [124] is NP-Complete, as it involves computing the inradii of $\mathcal{V}$-Polytopes [67]. When $\lambda \to \infty$, Theorem 4.1 reduces to the first computationally tractable guarantee for SSC that works for disjoint and potentially overlapping subspaces.

4.3.2 Randomized Results

We now present results for the random design case, i.e., data are generated under some random models.

Definition 4.5 (Random data)**.** ***"Random sampling"*** *assumes that for each ℓ, data points in $X^{(\ell)}$ are iid uniformly distributed on the unit sphere of $\mathcal{S}_\ell$.* ***"Random subspace"*** *assumes each $\mathcal{S}_\ell$ is generated independently by spanning d_ℓ iid uniformly distributed vectors on the unit sphere of $\mathbb{R}^n$.*

Lemma 4.1 (Singular value bound)**.** *Assume* random sampling. *If $d_\ell < N_\ell < n$, then there exists an absolute constant C_1 such that with probability of at least $1 - N_\ell^{-10}$,*

$$\sigma_{d_\ell}(X) \geq \frac{1}{2}\left(\sqrt{\frac{N_\ell}{d_\ell}} - 3 - C_1\sqrt{\frac{\log N_\ell}{d_\ell}}\right), \quad \textit{or simply} \quad \sigma_{d_\ell}(X) \geq \frac{1}{4}\sqrt{\frac{N_\ell}{d_\ell}},$$

if we assume $N_\ell \geq C_2 d_\ell$, for some constant C_2.

[1] Our simulation in Section 4.6 also supports this conjecture.

Lemma 4.2 (Inradius bound [2, 124])**.** *Assume* random sampling *of* $N_\ell = \kappa_\ell d_\ell$ *data points in each* $\mathcal{S}_\ell$*, then with probability larger than* $1 - \sum_{\ell=1}^{L} N_\ell e^{-\sqrt{d_\ell N_\ell}}$

$$r(\mathrm{conv}(\pm X_{-k}^{(\ell)})) \geq c(\kappa_\ell)\sqrt{\frac{\log(\kappa_\ell)}{2d_\ell}} \text{ for all pairs } (\ell, k).$$

Here, $c(\kappa_\ell)$ *is a constant depending on* κ_ℓ*. When* κ_ℓ *is sufficiently large, we can take* $c(\kappa_\ell) = 1/\sqrt{8}$*.*

Combining Lemma 4.1 and Lemma 4.2, we get the following remark showing that conditions (4.2) and (4.3) are complementary.

Remark 4.4. Under the *random sampling* assumption, when λ is smaller than a threshold, the singular value condition (4.2) is better than the inradius condition (4.3). Specifically, $\sigma_{d_\ell}(X) > \frac{1}{4}\sqrt{\frac{N_\ell}{d_\ell}}$ with high probability, so for some constant $C > 1$, the singular value condition is strictly better if

$$\lambda < \frac{C\left(\sqrt{N_\ell} - \sqrt{\log(N_\ell/d_\ell)}\right)}{\sqrt{N_\ell}\left(1 + \sqrt{\log(N_\ell/d_\ell)}\right)}, \quad \text{or when } N_\ell \text{ is large,} \quad \lambda < \frac{C}{1 + \sqrt{\log(N_\ell/d_\ell)}}.$$

By further assuming *random subspace*, we provide an upper bound of the incoherence μ.

Lemma 4.3 (Subspace incoherence bound)**.** *Assume* random subspace *and* random sampling*. It holds with probability greater than* $1 - 2/N$ *that for all* ℓ*,*

$$\mu(X^{(\ell)}) \leq \sqrt{\frac{6\log N}{n}}.$$

Combining Lemma 4.1 and Lemma 4.3, we have the following theorem.

Theorem 4.2 (LRSSC for random data)**.** *Suppose* L *rank-*d *subspace are uniformly and independently generated from* $\mathbb{R}^n$*, and* N/L *data points are uniformly and independently sampled from the unit sphere embedded in each subspace, furthermore* $N > CdL$ *for some absolute constant* C*, then SEP holds with probability larger than*

$1 - 2/N - 1/(Cd)^{10}$, *if*

$$d < \frac{n}{96 \log N}, \quad \textit{for all} \quad \lambda > \frac{1}{\sqrt{\frac{N}{L}}\left(\sqrt{\frac{n}{96 d \log N}} - 1\right)}. \tag{4.4}$$

The above condition is obtained from the singular value condition. Using the inradius guarantee, combined with Lemma 4.2 and 4.3, we have a different succeed condition requiring $d < \frac{n \log(\kappa)}{96 \log N}$ for all $\lambda > \frac{1}{\sqrt{\frac{n \log \kappa}{96 d \log N}} - 1}$. Ignoring constant terms, the condition on d is slightly better than (4.4) by a log factor but the range of valid λ is significantly reduced.

4.4 Graph Connectivity Problem

The graph connectivity problem concerns when SEP is satisfied, whether each block of the solution C to LRSSC represents a connected graph. The graph connectivity problem concerns whether each disjoint block (since SEP holds true) of the solution C to LRSSC represents a connected graph. This is equivalent to the connectivity of the solution of the following fictitious optimization problem, where each sample is constrained to be represented by the samples of the same subspace,

$$\min_{C^{(\ell)}} \|C^{(\ell)}\|_* + \lambda \|C^{(\ell)}\|_1 \quad s.t. \quad X^{(\ell)} = X^{(\ell)} C^{(\ell)}, \quad \operatorname{diag}(C^{(\ell)}) = 0. \tag{4.5}$$

The graph connectivity for SSC is studied by [104] under deterministic conditions (to make the problem well-posed). They show by a negative example that even if the well-posed condition is satisfied, the solution of SSC may not satisfy graph connectivity if the dimension of the subspace is greater than 3. On the other hand, graph connectivity problem is not an issue for LRR: as the following proposition suggests, the intra-class connections of LRR's solution are inherently dense (fully connected).

Proposition 4.1. *When the subspaces are independent, X is not full-rank and the data points are randomly sampled from a unit sphere in each subspace, then the solution to*

LRR, i.e.,

$$\min_C \|C\|_* \quad s.t. \quad X = XC,$$

is class-wise dense, namely each diagonal block of the matrix C is all non-zero.

The proof makes use of the following lemma which states the closed-form solution of LRR.

Lemma 4.4 ([98])**.** *Take skinny SVD of data matrix $X = U\Sigma V^T$. The closed-form solution to LRR is the shape interaction matrix $C = VV^T$.*

Proposition 4.1 then follows from the fact that each entry of VV^T has a continuous distribution, hence the probability that any is exactly zero is negligible (a complete argument is given in the Appendix).

Readers may notice that when $\lambda \to 0$, (4.5) is not exactly LRR, but with an additional constraint that diagonal entries are zero. We suspect this constrained version also have dense solution. This is demonstrated numerically in Section 4.6.

4.5 Practical issues

4.5.1 Data noise/sparse corruptions/outliers

The natural extension of LRSSC to handle noise is

$$\min_C \frac{1}{2}\|X - XC\|_F^2 + \beta_1\|C\|_* + \beta_2\|C\|_1 \quad s.t. \quad \operatorname{diag}(C) = 0. \tag{4.6}$$

We believe it is possible (but maybe tedious) to extend our guarantee to this noisy version following the strategy of [143] which analyzed the noisy version of SSC. This is left for future research.

According to the noisy analysis of SSC, a rule of thumb of choosing the scale of β_1 and β_2 is

$$\beta_1 = \frac{\sigma(\frac{1}{1+\lambda})}{\sqrt{2\log N}}, \qquad \beta_2 = \frac{\sigma(\frac{\lambda}{1+\lambda})}{\sqrt{2\log N}},$$

where λ is the tradeoff parameter used in noiseless case (4.1), σ is the estimated noise level and N is the total number of entries.

In case of sparse corruption, one may use ℓ_1 norm penalty instead of the Frobenious norm. For outliers, SSC is proven to be robust to them under mild assumptions [124], and we suspect a similar argument should hold for LRSSC too.

4.5.2 Fast Numerical Algorithm

As subspace clustering problem is usually large-scale, off-the-shelf SDP solvers are often too slow to use. Instead, we derive *alternating direction methods of multipliers* (ADMM) [17], known to be scalable, to solve the problem numerically. The algorithm involves separating out the two objectives and diagonal constraints with dummy variables C_2 and J like

$$\begin{aligned} &\min_{C_1,C_2,J} \ \|C_1\|_* + \lambda\|C_2\|_1 \\ &s.t. \quad X = XJ, \quad J = C_2 - \mathrm{diag}(C_2), \quad J = C_1, \end{aligned} \tag{4.7}$$

and update J, C_1, C_2 and the three dual variables alternatively. Thanks to the change of variables, all updates can be done in closed-form. To further speed up the convergence, we adopt the adaptive penalty mechanism of Lin et.al [94], which in some way ameliorates the problem of tuning numerical parameters in ADMM. Detailed derivations, update rules, convergence guarantee and the corresponding ADMM algorithm for the noisy version of LRSSC are made available in the appendix.

4.6 Numerical Experiments

To verify our theoretical results and illustrate the advantages of LRSSC, we design several numerical experiments. In all our numerical experiments, we use the ADMM implementation of LRSSC with fixed set of numerical parameters. The results are given against an exponential grid of λ values, so comparisons to only 1-norm (SSC) and only nuclear norm (LRR) are clear from two ends of the plots.

4.6.1 Separation-Sparsity Tradeoff

We first illustrate the tradeoff of the solution between obeying SEP and being connected (this is measured using the intra-class sparsity of the solution). We randomly generate L subspaces of dimension 10 from $\mathbb{R}^{50}$. Then, 50 unit length random samples are drawn from each subspace and we concatenate into a $50 \times 50L$ data matrix. We use Relative Violation [143] to measure of the violation of SEP and Gini Index [75] to measure the intra-class sparsity[1]. These quantities are defined below:

$$\text{RelViolation}\,(C, \mathcal{M}) = \frac{\sum_{(i,j)\notin\mathcal{M}} |C|_{i,j}}{\sum_{(i,j)\in\mathcal{M}} |C|_{i,j}},$$

where $\mathcal{M}$ is the index set that contains all (i, j) such that $x_i, x_j \in S_\ell$ for some ℓ.

$\text{GiniIndex}\,(C, \mathcal{M})$ is obtained by first sorting the absolute value of $C_{ij\in\mathcal{M}}$ into a non-decreasing sequence $\vec{c} = [c_1, ..., c_{|\mathcal{M}|}]$, then evaluate

$$\text{GiniIndex}\,(\text{vec}(C_{\mathcal{M}})) = 1 - 2\sum_{k=1}^{|\mathcal{M}|} \frac{c_k}{\|\vec{c}\|_1}\left(\frac{|\mathcal{M}| - k + 1/2}{|\mathcal{M}|}\right).$$

Note that RelViolation takes the value of $[0, \infty]$ and SEP is attained when RelViolation is zero. Similarly, Gini index takes its value in $[0, 1]$ and it is larger when intra-class connections are sparser.

The results for $L = 6$ and $L = 11$ are shown in Figure 4.1. We observe phase transitions for both metrics. When $\lambda = 0$ (corresponding to LRR), the solution does not obey SEP even when the independence assumption is only slightly violated ($L = 6$). When λ is greater than a threshold, RelViolation goes to zero. These observations match Theorems 4.1 and 4.2. On the other hand, when λ is large, intra-class sparsity is high, indicating possible disconnection within the class.

Moreover, we observe that there exists a range of λ where RelViolation reaches zero yet the sparsity level does not reaches its maximum. This justifies our claim that the

[1] We choose Gini Index over the typical ℓ_0 to measure sparsity as the latter is vulnerable to numerical inaccuracy.

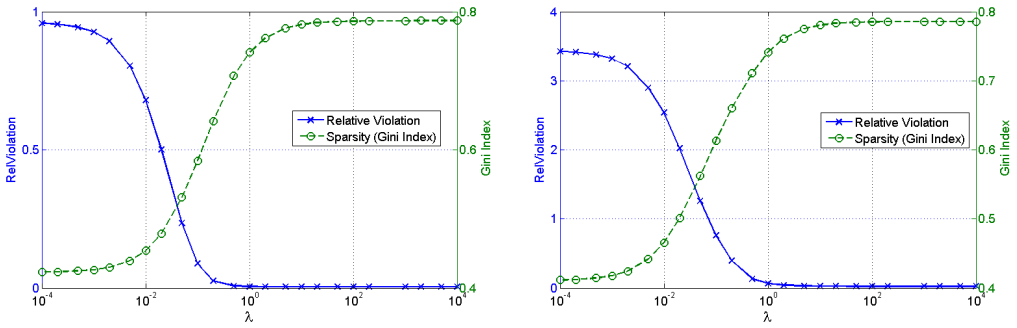


Figure 4.1: Illustration of the separation-sparsity trade-off. Left: 6 subspaces. Right: 11 subspace.

solution of LRSSC, taking λ within this range, can achieve SEP and at the same time keep the intra-class connections relatively dense. Indeed, for the subspace clustering task, a good tradeoff between separation and intra-class connection is important.

4.6.2 Skewed data distribution and model selection

In this experiment, we use the data for $L = 6$ and combine the first two subspaces into one 20-dimensional subspace and randomly sample 10 more points from the new subspace to "connect" the 100 points from the original two subspaces together. This is to simulate the situation when data distribution is skewed, i.e., the data samples within one subspace has two dominating directions. The skewed distribution creates trouble for model selection (judging the number of subspaces), and intuitively, the graph connectivity problem might occur.

We find that model selection heuristics such as the spectral gap [141] and spectral gap ratio [90] of the normalized Laplacian are good metrics to evaluate the quality of the solution of LRSSC. Here the correct number of subspaces is 5, so the spectral gap is the difference between the 6^{th} and 5^{th} smallest singular value and the spectral gap ratio is the ratio of adjacent spectral gaps. The larger these quantities, the better the affinity matrix reveals that the data contains 5 subspaces.

Figure 4.2 demonstrates how singular values change when λ increases. When $\lambda = 0$ (corresponding to LRR), there is no significant drop from the 6^{th} to the 5^{th} singular

value, hence it is impossible for either heuristic to identify the correct model. As λ increases, the last 5 singular values gets smaller and become almost zero when λ is large. Then the 5-subspace model can be correctly identified using spectral gap ratio. On the other hand, we note that the 6^{th} singular value also shrinks as λ increases, which makes the spectral gap very small on the SSC side and leaves little robust margin for correct model selection against some violation of SEP. As is shown in Figure 4.3, the largest spectral gap and spectral gap ratio appear at around $\lambda = 0.1$, where the solution is able to benefit from both the better separation induced by the 1-norm factor and the relatively denser connections promoted by the nuclear norm factor.

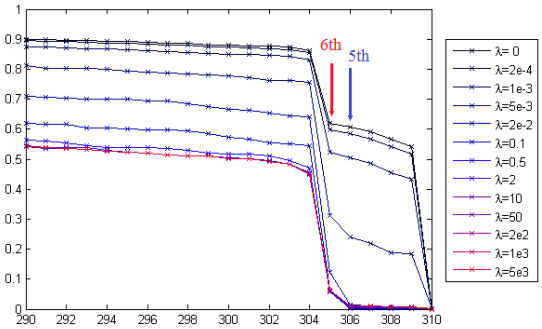


Figure 4.2: Last 20 singular values of the normalized Laplacian in the skewed data experiment.

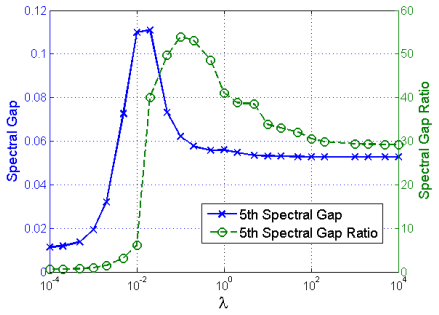


Figure 4.3: Spectral Gap and Spectral Gap Ratio in the skewed data experiment.

4.7 Additional experimental results

4.7.1 Numerical Simulation

Exp1: Disjoint 11 Subspaces Experiment

Randomly generate 11 subspaces of dimension 10 from $\mathbb{R}^{50}$. 50 unit length random samples are drawn from each subspace and we concatenate into a 50×550 data matrix. Besides what is shown in the main text, we provide a qualitative illustration of the separation-sparsity trade-off in Figure 4.4.

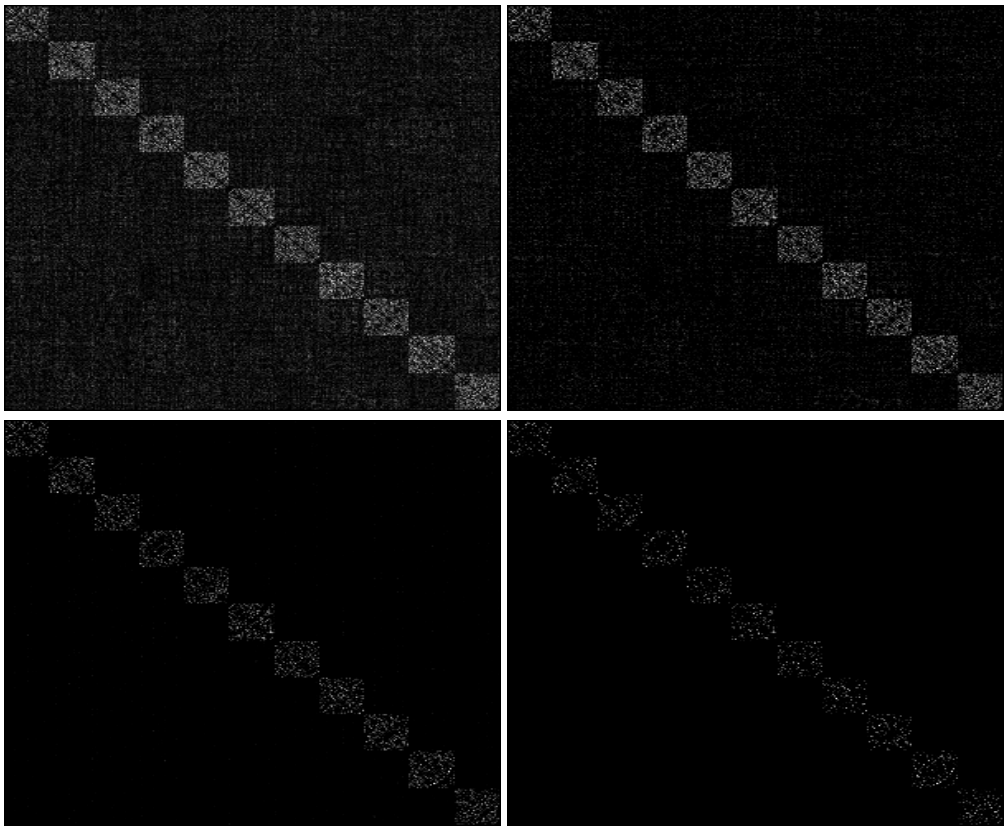


Figure 4.4: Qualitative illustration of the 11 Subspace Experiment. From left to right, top to bottom: $\lambda = [0, 0.05, 1, 1e4]$, corresponding RelViolation is [3.4, 1.25, 0.06, 0.03] and Gini Index is [0.41, 0.56, 0.74, 0.79]

Exp2: when exact SEP is not possible

In this experiment, we randomly generate 10 subspaces of rank 3 from a 10 dimensional subspace, each sampled 15 data points. All data points are embedded to the ambient space of dimension 50.

This is to illustrate the case when perfect SEP is not possible for any λ. In other word, the smallest few singular values of the normalized Laplacian matrix is not exactly 0. Hence we will rely on heuristics such as Spectral Gap and Spectral Gap Ratio to tell how many subspaces there are and hopefully spectral clustering will return a good clustering. Figure 4.5 gives an qualitative illustration how the spectral gap emerges as λ increases. Figure 4.6 shows quantitatively the same thing with the actual values of the two heuristics changes. Clearly, model selection is much easier in the SSC-side comparing to the LRR side, when SEP is the main issue (see the comparison in Figure 4.7).

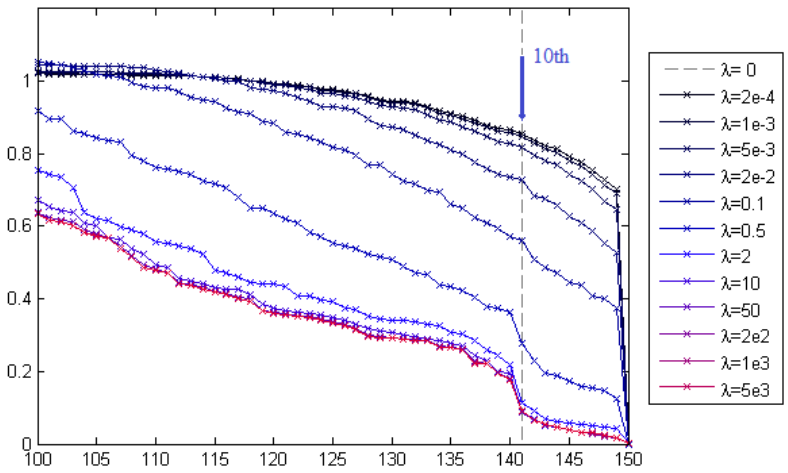


Figure 4.5: Last 50 Singular values of the normalized Laplacian in Exp2. See how the spectral gap emerges and become larger as λ increases.

Exp3: Independent-Skewed data distribution

Assume ambient dimension $n = 50$, 3 subspaces. The second and the third 3-d subspaces are generated randomly, each sampled 15 points. The first subspace is a 6-d

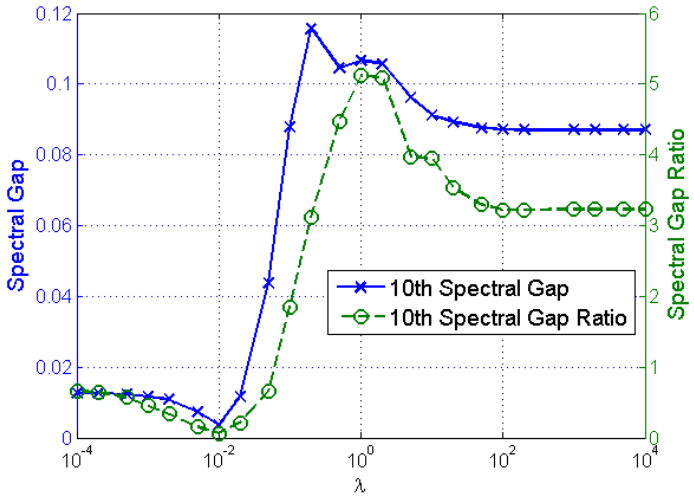


Figure 4.6: Spectral Gap and Spectral Gap Ratio for Exp2. When perfect SEP is not possible, model selection is easier on the SSC side, but the optimal spot is still somewhere between LRR and SSC.

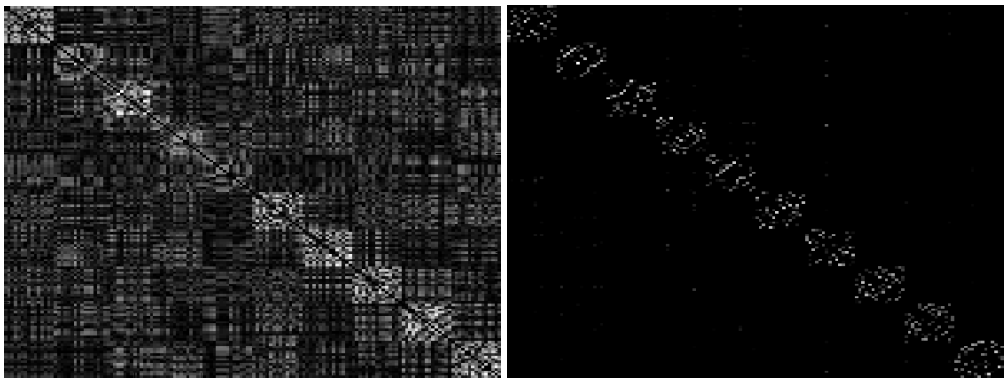


Figure 4.7: Illustration of representation matrices. Left: $\lambda = 0$, Right: $\lambda = 1e4$. While it is still not SEP, there is significant improvement in separation.

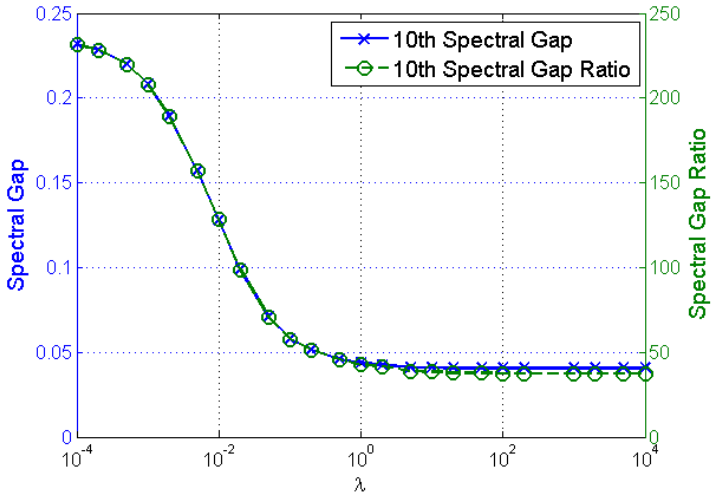


Figure 4.8: Spectral Gap and Spectral Gap Ratio for Exp3. The independent subspaces have no separation problem, SEP holds for all λ. Note that due to the skewed data distribution, the spectral gap gets quite really small at the SSC side.

subspace spanned by two random 3-d subspaces. 15 data points are randomly generated from each of the two spanning 3-d subspaces and only 3 data points are randomly taken from the spanned 6-D subspace two glue them together.

As a indication of model selection, the spectral gap and spectral ratio for all λ is shown in Figure 4.8. While all experiments return clearly defined three disjoint components (smallest three singular values equal to 0 for all λ), the LRR side gives the largest margin of three subspaces (when $\lambda = 0$, the result gives the largest 4th smallest singular value). This illustrates that when Skewed-Data-Distribution is the main issue, LRR side is better than SSC side. This can be qualitatively seen in Figure 4.9

Exp4: Disjoint-Skewed data distribution

In this experiment, we illustrate the situation when subspaces are not independent and one of them has skewed distribution, hence both LRR and SSC are likely to to encounter problems. The setup is the same as the 6 Subspace experiment except the first two subspaces are combined into a 20-dimensional subspace moreover 10 more random points are sampled from the spanned subspace. Indeed, as Figure 4.2 and 4.3 suggest,

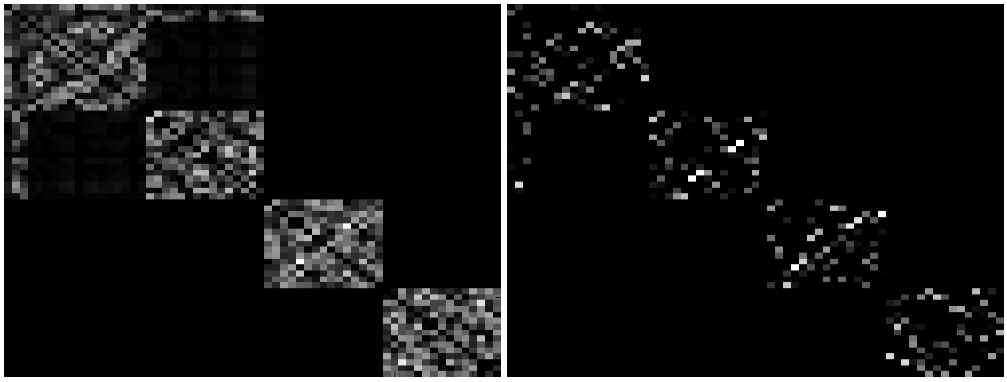


Figure 4.9: Illustration of representation matrices. Left: $\lambda = 0$, Right: $\lambda = 1e4$. The 3 diagonal block is clear on the LRR side, while on the SSC side, it appear to be more like 4 blocks plus some noise.

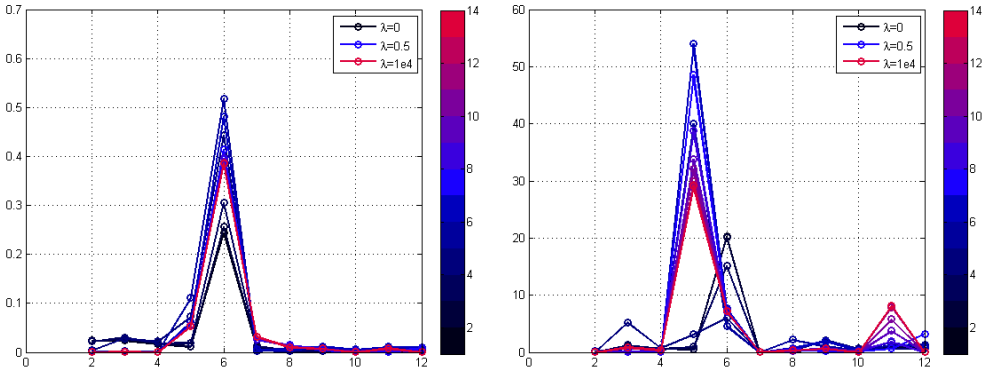


Figure 4.10: Illustration of model selection with spectral gap (left) and spectral gap ratio (right) heuristic. The highest point of each curve corresponds to the inferred number of subspaces in the data. We know the true number of subspace is 5.

taking λ somewhere in the middle gives the largest spectral gap and spectral gap ratio, which indicates with large margin that the correct model is a 5 Subspace Model.

In addition to that, we add Figure 4.10 here to illustrate the ranges of λ where two heuristics give correct model selection. It appears that "spectral gap" suggests a wrong model for all λ despite the fact that the 5^{th} "spectral gap" enlarges as λ increase. On the other hand, the "spectral gap ratio" reverted its wrong model selection at the LRR side quickly as λ increases and reaches maximum margin in the blue region (around $\lambda = 0.5$). This seems to imply that "spectral gap ratio" is a better heuristic in the case when one or more subspaces are not well-represented.

4.7.2 Real Experiments on Hopkins155

To complement the numerical experiments, we also run our NoisyLRSSC on the Hopkins155 motion segmentation dataset[136]. The dataset contains 155 short video sequence with temporal trajectories of the 2D coordinates of the feature points summarizing in a data matrix. The task is to unsupervisedly cluster the given trajectories into blocks such that each block corresponds to one rigid moving objects. The motion can be 3D translation, rotation or combination of translation and rotation. Ground truth is given together with the data so evaluation is simply by the misclassification rate. A few snapshots of the dataset is given in Figure 4.11.

4.7.2.1 Why subspace clustering?

Subspace clustering is applicable here because collections of feature trajectories on a rigid body captured by a moving affine camera can be factorized into camera motion matrix and a structure matrix as follows

$$X = \begin{pmatrix} x_{11} & \dots & x_{1n} \\ \dots & \dots & \dots \\ x_{m1} & \dots & x_{mn} \end{pmatrix} = \begin{pmatrix} M_1 \\ \dots \\ M_m \end{pmatrix} \begin{pmatrix} S_1 & \dots & S_n \end{pmatrix},$$

where $M_i \in \mathbb{R}^{2\times 4}$ is a the camera projection matrix from 3D homogeneous coordinates to 2D image coordinates and $S_j \in \mathbb{R}^4$ is one feature points in 3D with 1 added at the back to form the homogeneous coordinates. Therefore, the inner dimension of the matrix multiplication ensures that all column vectors of X lies in a 4 dimensional subspace (see [69, Chapter 18] for details).

Depending on the types of motion, and potential projective distortion of the image (real camera is never perfectly affine) the subspace may be less than rank 4 (degenerate motion) or only approximately rank 4.

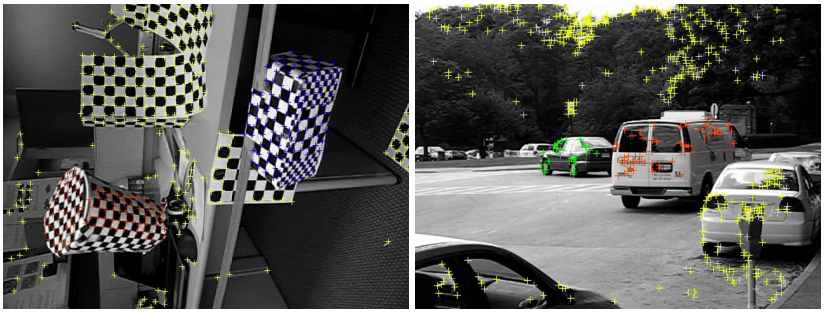


Figure 4.11: Snapshots of Hopkins155 motion segmentation data set.

4.7.2.2 Methods

We run the ADMM version of the NoisyLRSSC (C.22) using the same parameter scheme (but with different values) proposed in [57] for running Hopkins155. Specifically, we rescaled the original problem into:

$$\begin{aligned} &\min_{C_1,C_2,J} \frac{\alpha}{2}\|X - XJ\|_F^2 + \alpha\beta_1\|C_1\|_* + \alpha\beta_2\|C_2\|_1 \\ &s.t. \quad J = C_2 - \operatorname{diag}(C_2), \quad J = C_1, \end{aligned}$$

and set

$$\alpha = \frac{\alpha_z}{\mu_z}, \qquad \beta_1 = \frac{1}{1+\lambda}, \qquad \beta_2 = \frac{\lambda}{1+\lambda}.$$

with $\alpha_z = 15000$[1], and

$$\mu_z = \min_i \max_{i \neq j} \langle x_i, x_j \rangle.$$

Numerical parameters in the Lagrangian are set to $\mu_2 = \mu_3 = 0.1\alpha$. Note that we have a simple adaptive parameter that remains constant for each data sequence.

Also note that we do not intend to tune the parameters to its optimal and outperform the state-of-the-art. This is just a minimal set of experiments on the real data to justify how the combinations of the two objectives may be useful when all other factors are equal.

[1] In [57], they use $\alpha_z = 800$, but we find it doesn't work out in our case. We will describe the difference to their experiments on Hopkins155 separately later.

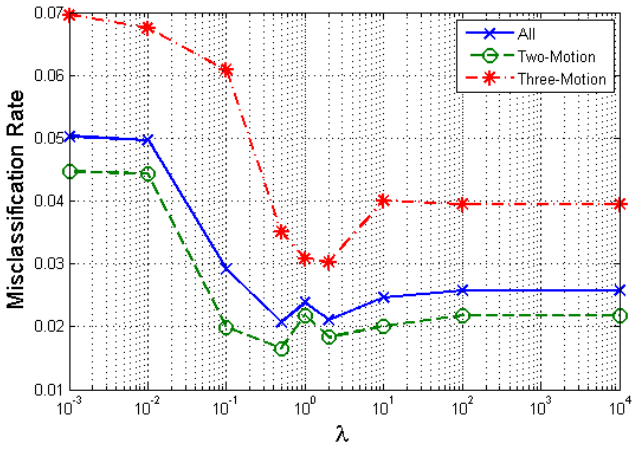


Figure 4.12: Average misclassification rates vs. λ.

4.7.2.3 Results

Figure 4.12 plots how average misclassification rate changes with λ. While it is not clear on the two-motion sequences, the advantage of LRSSC is drastic on three motions.

To see it more clearly, we plot the RelViolation, Gini index and misclassification of all sequence for all λ in Figure 4.14, Figure 4.15 and Figure 4.13 respectively. From Figure 4.14 and 4.15, we can tell that the shape is well predicted by our theorem and simulation. Since a correct clustering depends on both inter-class separation and intra-class connections, it is understandable that we observe the phenomena in Figure 4.13 that some sequences attain zero misclassification on the LRR side, some on the SSC side, and to our delight, some reaches the minimum misclassification rate somewhere in between.

4.7.2.4 Comparison to SSC results in [57]

After carefully studying the released SSC code that generates Table 5 in [57], we realized that they use two post processing steps on the representation matrix C before constructing affinity matrix $|C| + |C^T|$ for spectral clustering. First, they use a thresholding step to keep only the largest non-zero entries that sum to 70% of the ℓ_1 norm of each column. Secondly, there is a normalization step that scales the largest entry in

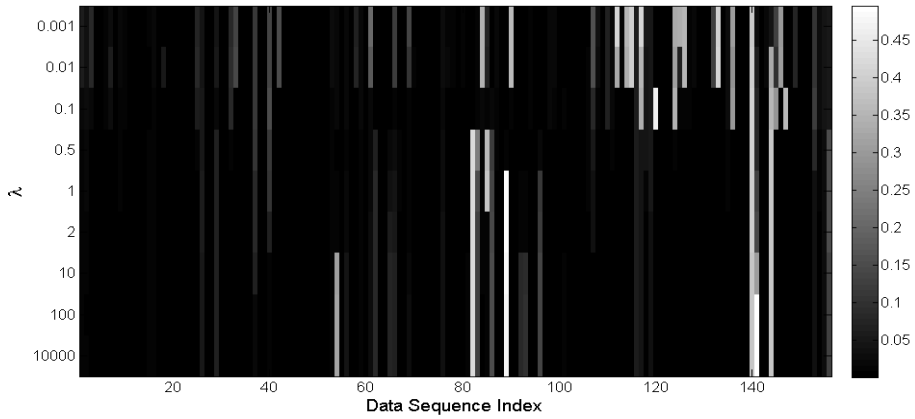


Figure 4.13: Misclassification rate of the 155 data sequence against λ. Black regions refer to perfect clustering, and white regions stand for errors.

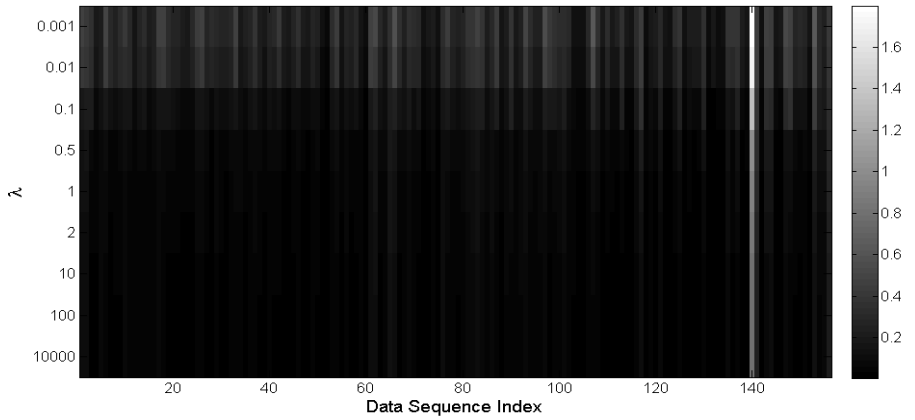


Figure 4.14: RelViolation of representation matrix C in the 155 data sequence against λ. Black regions refer to zero RelViolation (namely, SEP), and white regions stand for large violation of SEP.

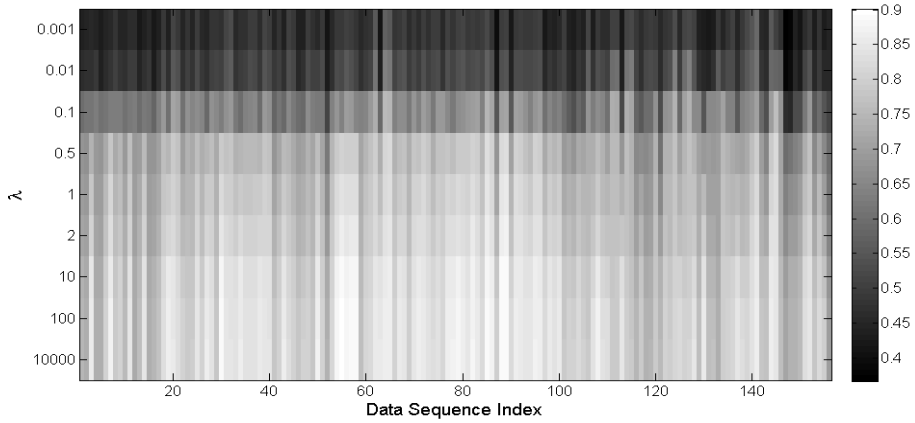


Figure 4.15: GiniIndex of representation matrix C in the 155 data sequence against λ. Darker regions represents denser intra-class connections, lighter region means that the connections are sparser.

each column to one (and the rest accordingly). The results with 4.4% and 1.95% misclassification rates for respectively 3-motion and 2-motion sequences essentially refer to the results with postprocessing.

Without postprocessing, the results we get are 5.67% for 3-motions and 1.91% for 2-motions. Due to the different implementation of the numerical algorithms (in stopping conditions and etc), we are unable to reproduce the same results on the SSC end (when λ is large) with the same set of weighting factor, but we managed to make the results comparable (slightly better) with a different set of weighting even without any post-processing steps. Moreover, when we choose λ such that we have a meaningful combination of ℓ_1 norm and nuclear norm regularization, the 3-motion misclassification rate goes down to 3%.

Since the Hopkins155 dataset is approaching saturation, it is not our point to conclude that a few percentage of improvement is statistically meaningful, since one single failure case that has 40% of misclassification will already raise the overall misclassification rate by 1.5%. Nevertheless, we are delighted to see LRSSC in its generic form performs in a comparable level as other state-of-the-art algorithms.

4.8 Chapter Summary

In this chapter, we proposed LRSSC for the subspace clustering problem and provided theoretical analysis of the method. We demonstrated that LRSSC is able to achieve perfect SEP for a wider range of problems than previously known for SSC and meanwhile maintains denser intra-class connections than SSC (hence less likely to encounter the "graph connectivity" issue). Furthermore, the results offer new understandings to SSC and LRR themselves as well as problems such as skewed data distribution and model selection. An important future research question is to mathematically define the concept of the graph connectivity, and establish conditions that perfect SEP and connectivity indeed occur together for some non-empty range of λ for LRSSC.

Chapter 5

PARSuMi: Practical Matrix Completion and Corruption Recovery with Explicit Modeling

Low-rank matrix completion is a problem of immense practical importance. Recent works on the subject often use nuclear norm as a convex surrogate of the rank function. Despite its solid theoretical foundation, the convex version of the problem often fails to work satisfactorily in real-life applications. Real data often suffer from very few observations, with support not meeting the random requirements, ubiquitous presence of noise and potentially gross corruptions, sometimes with these simultaneously occurring. This chapter proposes a Proximal Alternating Robust Subspace Minimization (PARSuMi) method to tackle the three problems. The proximal alternating scheme explicitly exploits the rank constraint on the completed matrix and uses the ℓ_0 pseudo-norm directly in the corruption recovery step. We show that the proposed method for the non-convex and non-smooth model converges to a stationary point. Although it is not guaranteed to find the global optimal solution, in practice we find that our algorithm can typically arrive at a good local minimizer when it is supplied with a reasonably good starting point based on convex optimization. Extensive experiments with challenging

synthetic and real data demonstrate that our algorithm succeeds in a much larger range of practical problems where convex optimization fails, and it also outperforms various state-of-the-art algorithms. Part of the materials in this chapter is included in our manuscript [144] that is currently under review.

5.1 Introduction

Completing a low-rank matrix from partially observed entries, also known as matrix completion, is a central task in many real-life applications. The same abstraction of this problem has appeared in diverse fields such as signal processing, communications, information retrieval, machine learning and computer vision. For instance, the missing data to be filled in may correspond to plausible movie recommendations [61, 87], occluded feature trajectories for rigid or non-rigid structure from motion, namely SfM [19, 68] and NRSfM [111], relative distances of wireless sensors [107], pieces of uncollected measurements in DNA micro-array [60], just to name a few.

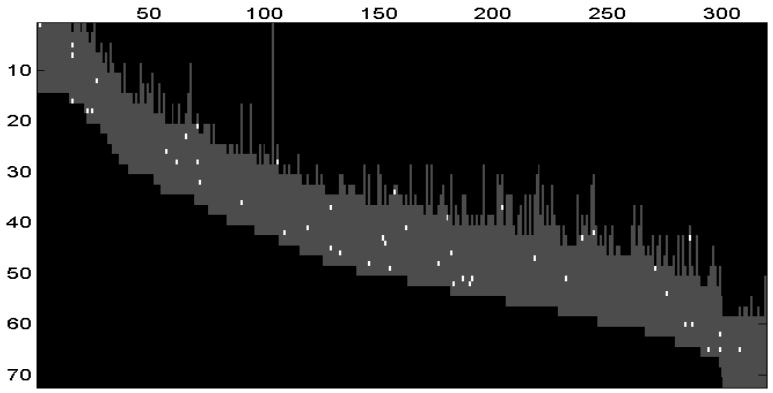


Figure 5.1: Sampling pattern of the Dinosaur sequence: 316 features are tracked over 36 frames. Dark area represents locations where no data is available; sparse highlights are injected gross corruptions. Middle stripe in grey are noisy observed data, occupying 23% of the full matrix. The task of this chapter is to fill in the missing data and recover the corruptions.

The common difficulty of these applications lies in the scarcity of the observed data, uneven distribution of the support, noise, and more often than not, the presence of gross

corruptions in some observed entries. For instance, in the movie rating database Netflix [14], only less than 1% of the entries are observed and 90% of the observed entries correspond to 10% of the most popular movies. In photometric stereo, the missing data and corruptions (arising from shadow and specular highlight as modeled in Wu et al. [149]) form contiguous blocks in images and are by no means random. In structure from motion, the observations fall into a diagonal band shape, and feature coordinates are often contaminated by tracking errors (see the illustration in Figure 5.1). Therefore, in order for any matrix completion algorithm to work in practice, these aforementioned difficulties need to be tackled altogether. We refer to this problem as **practical matrix completion**. Mathematically, the problem to be solved is the following:

Given	$\Omega, \widehat{W}_{ij}$ for all $(i,j) \in \Omega$,
find	$W, \tilde{\Omega}$,
s.t.	$\mathrm{rank}(W)$ is small; $\mathrm{card}(\tilde{\Omega})$ is small;
	$\lvert W_{ij} - \widehat{W}_{ij} \rvert$ is small $\forall (i,j) \in \Omega \lvert \tilde{\Omega}$.

where Ω is the index set of observed entries whose locations are not necessarily selected at random, $\tilde{\Omega} \in \Omega$ represents the index set of corrupted data, $\widehat{W} \in \mathbb{R}^{m \times n}$ is the measurement matrix with only $\widehat{W}_{ij \in \Omega}$ known, i.e., its support is contained in Ω. Furthermore, we define the projection $\mathcal{P}_\Omega : \mathbb{R}^{m \times n} \mapsto \mathbb{R}^{|\Omega|}$ so that $\mathcal{P}_\Omega(\widehat{W})$ denotes the vector of observed data. The adjoint of $\mathcal{P}_\Omega$ is denoted by $\mathcal{P}^*_\Omega$.

Extensive theories and algorithms have been developed to tackle some aspect of the challenges listed in the preceding paragraph, but those tackling the full set of challenges are far and few between, thus resulting in a dearth of practical algorithms. Two dominant classes of approaches are nuclear norm minimization, e.g. Candes and Plan [21], Candes and Recht [24], Candès et al. [27], Chen et al. [39], and matrix factorization, e.g., Buchanan and Fitzgibbon [19], Chen [36], Eriksson and Van Den Hengel [58], Koren et al. [87], Okatani and Deguchi [108]. Nuclear norm minimization methods minimize the convex relaxation of rank instead of the rank itself, and are supported

by rigorous theoretical analysis and efficient numerical computation. However, the conditions under which they succeed are often too restrictive for it to work well in real-life applications (as reported in Shi and Yu [119] and Jain et al. [76]). In contrast, matrix factorization is widely used in practice and are considered very effective for problems such as movie recommendation [87] and structure from motion [111, 135] despite its lack of rigorous theoretical foundation. Indeed, as one factorizes matrix W into UV^T, the formulation becomes bilinear and thus optimal solution is hard to obtain except in very specific cases (e.g., in Jain et al. [76]). A more comprehensive survey of the algorithms and review of the strengths and weaknesses will be given in the next section.

In this chapter, we attempt to solve the practical matrix completion problem under the prevalent case where the rank of the matrix W and the cardinality of $\tilde{\Omega}$ are upper bounded by some known parameters r and N_0 via the following non-convex, non-smooth optimization model:

$$\begin{aligned} \min_{W,E} \quad & \tfrac{1}{2}\|\mathcal{P}_{\Omega}(W - \widehat{W} + E)\|^2 + \tfrac{\lambda}{2}\|\mathcal{P}_{\overline{\Omega}}(W)\|^2 \\ \text{s.t.} \quad & \operatorname{rank}(W) \leq r,\ W \in \mathbb{R}^{m\times n} \\ & \|E\|_0 \leq N_0,\ \|E\| \leq K_E,\ E \in \mathbb{R}_{\Omega}^{m\times n} \end{aligned} \tag{5.1}$$

where $\mathbb{R}_{\Omega}^{m\times n}$ denotes the set of $m \times n$ matrices whose supports are subsets of Ω and $\|\cdot\|$ is the Frobenius norm; K_E is a finite constant introduced to facilitate the convergence proof. Note that the restriction of E to $\mathbb{R}_{\Omega}^{m\times n}$ is natural since the role of E is to capture the gross corruptions in the observed data $\widehat{W}_{ij\in\Omega}$. The bound constraint on E is natural in some problems when the true matrix W is bounded (e.g., Given the typical movie ratings of 0-10, the gross outliers can only lie in [-10, 10]). In other problems, we simply choose K_E to be some large multiple (say 20) of $\sqrt{N_0} \times \operatorname{median}(\mathcal{P}_{\Omega}(\widehat{W}))$, so that the constraint is essentially inactive and has no impact on the optimization. Note that without making any randomness assumption on the index set Ω or assuming that the problem has a unique solution (W^*, E^*) such that the singular vector matrices of W^* satisfy some inherent conditions like those in Candès et al. [27], the problem of

practical matrix completion is generally ill-posed. This motivated us to include the Tikhonov regularization term $\frac{\lambda}{2}\|\mathcal{P}_{\overline{\Omega}}(W)\|^2$ in (5.1), where $\overline{\Omega}$ denotes the complement of Ω, and $0 < \lambda < 1$ is a small constant. Roughly speaking, what the regularization term does is to pick the solution W which has the smallest $\|\mathcal{P}_{\overline{\Omega}}(W)\|$ among all the candidates in the optimal solution set of the non-regularized problem. Notice that we only put a regularization on those elements of W in $\overline{\Omega}$ as we do not wish to perturb those elements of W in the fitting term. Finally, with the Tikhonov regularization and the bound constraint on $\|E\|$, we can show that problem (5.1) has a global minimizer.

By defining $H \in \mathbb{R}^{m\times n}$ to be the matrix such that

$$H_{ij} = \begin{cases} 1 & \text{if } (i,j) \in \Omega \\ \sqrt{\lambda} & \text{if } (i,j) \notin \Omega, \end{cases} \tag{5.2}$$

we can rewrite the objective function in (5.1) in a compact form, and the problem becomes:

$$\begin{aligned} \min_{W,E} \quad & \tfrac{1}{2}\|H \circ (W + E - \widehat{W})\|^2 \\ \text{s.t.} \quad & \operatorname{rank}(W) \le r,\ W \in \mathbb{R}^{m\times n} \\ & \|E\|_0 \le N_0,\ \|E\| \le K_E,\ E \in \mathbb{R}^{m\times n}_{\Omega}. \end{aligned} \tag{5.3}$$

In the above, the notation "$\circ$" denotes the element-wise product between two matrices.

We propose PARSuMi, a proximal alternating minimization algorithm motivated by the algorithm in Attouch et al. [3] to solve (5.3). This involves solving two subproblems each with an auxiliary proximal regularization term. It is important to emphasize that the subproblems in our case are non-convex and hence it is essential to design appropriate algorithms to solve the subproblems to global optimality, at least empirically. We develop essential reformulations of the subproblems and design novel techniques to efficiently solve each subproblem, provably achieving the global optimum for one, and empirically so for the other. We also prove that our algorithm is guaranteed to converge to a limit point, which is necessarily a stationary point of (5.3). Together with

the initialization schemes we have designed based on the convex relaxation of (5.3), our method is able to solve challenging real matrix completion problems with corruptions robustly and accurately. As we demonstrate in the experiments, PARSuMi is able to provide excellent reconstruction of unobserved feature trajectories in the classic Oxford Dinosaur sequence for SfM, despite structured (as opposed to random) observation pattern and data corruptions. It is also able to solve photometric stereo to high precision despite severe violations of the Lambertian model (which underlies the rank-3 factorization) due to shadow, highlight and facial expression difference. Compared to state-of-the-art methods such as GRASTA [71], Wiberg ℓ_1 [58] and BALM [46], our results are substantially better both qualitatively and quantitatively.

Note that in (5.3) we do not seek convex relaxation of any form, but rather constrain the rank and the corrupted entries' cardinality directly in their original forms. While it is generally not possible to have an algorithm guaranteed to compute the global optimal solution, we demonstrate that with appropriate initializations, the faithful representation of the original problem often offers significant advantage over the convex relaxation approach in denoising and corruption recovery, and is thus more successful in solving real problems.

The rest of the chapter is organized as follows. In Section 5.2, we provide a comprehensive review of the existing theories and algorithms for practical matrix completion, summarizing the strengths and weaknesses of nuclear norm minimization and matrix factorization. In Section 5.3, we conduct numerical evaluations of predominant matrix factorization methods, revealing those algorithms that are less-likely to be trapped at local minima. Specifically, these features include parameterization on a subspace and second-order Newton-like iterations. Building upon these findings, we develop the PARSuMi scheme in Section 5.4 to simultaneously handle sparse corruptions, dense noise and missing data. The proof of convergence and a convex initialization scheme are also provided in this section. In Section 5.5, the proposed method is evaluated on both synthetic and real data and is shown to outperform the current state-of-the-art algorithms for robust matrix completion.

5.2 A survey of results

5.2.1 Matrix completion and corruption recovery via nuclear norm minimization

	MC[32]	RPCA [27]	NoisyMC [22]	StableRPCA [155]	RMC[93]	RMC[39]
Missing data	Yes	No	Yes	No	Yes	Yes
Corruptions	No	Yes	No	Yes	Yes	Yes
Noise	No	No	Yes	Yes	No	No
Deterministic Ω	No	No	No	No	No	Yes
Deterministic $\tilde{\Omega}$	No	No	No	No	No	Yes

Table 5.1: Summary of the theoretical development for matrix completion and corruption recovery.

Recently, the most prominent approach for solving a matrix completion problem is via the following nuclear norm minimization:

$$\min_{W} \left\{ \|W\|_* \,\middle|\, \mathcal{P}_\Omega(W - \widehat{W}) = 0 \right\}, \tag{5.4}$$

in which $\mathrm{rank}(X)$ is replaced by the nuclear norm $\|X\|_* = \sum_i \sigma_i(X)$, where the latter is the tightest convex relaxation of rank over the unit (spectral norm) ball. Candes and Recht [24] showed that when sampling is uniformly random and sufficiently dense, and the underlying low-rank subspace is *incoherent* with respect to the standard bases, then the remaining entries of the matrix can be exactly recovered. The guarantee was later improved in Candès and Tao [29], Recht [114] and extended for noisy data in Candes and Plan [21], Negahban and Wainwright [105] relaxed the equality constraint to

$$\|\mathcal{P}_\Omega(W - \widehat{W})\| \leq \delta.$$

Using similar assumptions and arguments, Candès et al. [27] proposed solution to the related problem of robust principal component analysis (RPCA) where the low-rank matrix can be recovered from sparse corruptions (with no missing data). This is formu-

lated as

$$\min_{W,E} \left\{ \|W\|_* + \lambda \|E\|_1 \,\middle|\, W + E = \widehat{W} \right\}. \tag{5.5}$$

Using deterministic geometric conditions concerning the tangent spaces of the ground truth $(\bar{W}, \bar{E})$, Chandrasekaran et al. [34] also established strong recovery result via the convex optimization problem (5.5). Noisy extension and improvement of the guarantee for RPCA were provided by Zhou et al. [155] and Ganesh et al. [63] respectively. Chen et al. [39] and Li [93] combined (5.4) and (5.5) and provided guarantee for the following

$$\min_{W,E} \left\{ \|W\|_* + \lambda \|E\|_1 \,\middle|\, \mathcal{P}_\Omega(W + E - \widehat{W}) = 0 \right\}. \tag{5.6}$$

In particular, the results in Chen et al. [39] lifted the uniform random support assumptions in previous works by laying out the exact recovery condition for a class of deterministic sampling (Ω) and corruptions ($\tilde{\Omega}$) patterns.

We summarize the theoretical and algorithmic progress in practical matrix completion achieved by each method in Table 5.1. It appears that researchers are moving towards analyzing all possible combinations of the problems; from past indication, it seems entirely plausible albeit tedious to show the noisy extension

$$\min_{W,E} \left\{ \|W\|_* + \lambda \|E\|_1 \,\middle|\, \|\mathcal{P}_\Omega(W + E - \widehat{W})\| \leq \delta \right\} \tag{5.7}$$

will return a solution stable around the desired W and E under appropriate assumptions. Wouldn't that solve the practical matrix completion problem altogether?

The answer is unfortunately no. While this line of research have provided profound understanding of practical matrix completion itself, the actual performance of the convex surrogate on real problems (e.g., movie recommendation) is usually not competitive against nonconvex approaches such as matrix factorization. Although convex relaxation is amazingly equivalent to the original problem under certain conditions, those well versed in practical problems will know that those theoretical conditions are usually not satisfied by real data. Due to noise and model errors, real data are seldom truly

low-rank (see the comments on Jester joke dataset in Keshavan et al. [81]), nor are they as incoherent as randomly generated data. More importantly, observations are often structured (e.g., diagonal band shape in SfM) and hence do not satisfy the random sampling assumption needed for the tight convex relaxation approach. As a consequence of all these factors, the recovered W and E by convex optimization are often neither low-rank nor sparse in practical matrix completion. This can be further explained by the so-called "Robin Hood" attribute of ℓ_1 norm (analogously, nuclear norm is the ℓ_1 norm in the spectral domain), that is, it tends to steal from the rich and give it to the poor, decreasing the inequity of "wealth" distribution. Illustrations of the attribute will be given in Section 5.5.

Nevertheless, the convex relaxation approach has the advantage that one can design *efficient* algorithms to find or approximately reach the *global* optimal solution of the given convex formulation. In this chapter, we take advantage of the convex relaxation approach and use it to provide a powerful initialization for our algorithm to converge to the correct solution.

5.2.2 Matrix factorization and applications

Another widely-used method to estimate missing data in a low-rank matrix is matrix factorization (MF). It is at first considered as a special case of the weighted low-rank approximation problem with $\{0, 1\}$ weight by Gabriel and Zamir in 1979 and much later by Srebro and Jaakkola [127]. The buzz of Netflix Prize further popularizes the missing data problem as a standalone topic of research. Matrix factorization turns out to be a robust and efficient realization of the idea that people's preferences of movies are influenced by a small number of latent factors and has been used as a key component in almost all top-performing recommendation systems [87] including BellKor's Pragmatic Chaos, the winner of the Netflix Prize [86].

In computer vision, matrix factorization with missing data is recognized as an important problem too. Tomasi-Kanade affine factorization [135], Sturm-Triggs projective factorization [131], and many techniques in Non-Rigid SfM and motion tracking [111]

can all be formulated as a matrix factorization problem. Missing data and corruptions emerge naturally due to occlusions and tracking errors. For a more exhaustive survey of computer vision problems that can be modelled by matrix factorization, we refer readers to Del Bue et al. [46].

Regardless of its applications, the key idea is that when $W = UV^T$, one ensures that the required rank constraint is satisfied by restricting the factors U and V to be in $\mathbb{R}^{m\times r}$ and $\mathbb{R}^{n\times r}$ respectively. Since the (U, V) parameterization has a much smaller degree of freedom than the dimension of W, completing the missing data becomes a better posed problem. This gives rise to the following optimization problem:

$$\min_{U,V} \quad \frac{1}{2}\left\|\mathcal{P}_\Omega(UV^T - \widehat{W})\right\|^2 \tag{5.8}$$

or its equivalence reformulation

$$\min_{U} \left\{ \frac{1}{2}\left\|\mathcal{P}_\Omega(UV(U)^T - \widehat{W})\right\|^2 \middle| U^TU = I_r \right\} \tag{5.9}$$

where the factor V is now a function of U.

Unfortunately, (5.8) is not a convex optimization problem. The quality of the solutions one may get by minimizing this objective function depends on specific algorithms and their initializations. Roughly speaking, the various algorithms for (5.8) may be grouped into three categories: **alternating minimization**, **first order** gradient methods and **second order** Newton-like methods.

Simple approaches like alternating least squares (ALS) or equivalently PowerFactorization [68] fall into the first category. They alternatingly fix one factor and minimize the objective over the other using least squares method. A more sophisticated algorithm is BALM [46], which uses the Augmented Lagrange Multiplier method to gradually impose additional problem-specific manifold constraints. The inner loop however is still alternating minimization. This category of methods has the reputation of reducing the objective value quickly in the first few iterations, but they usually take a large number of iterations to converge to a high quality solution [19].

First order gradient methods are efficient, easy to implement and they are able to scale up to million-by-million matrices if stochastic gradient descent is adopted. Therefore it is very popular for large-scale recommendation systems. Typical approaches include Simon Funk's incremental SVD [61], nonlinear conjugate gradient [127] and more sophisticatedly, gradient descent on the Grassmannian/Stiefel manifold, such as GROUSE [7] and OptManifold [147]. These methods, however, as we will demonstrate later, easily get stuck in local minima[1].

The best performing class of methods are the second order Newton-like algorithms, in that they demonstrate superior performance in both accuracy and the speed of convergence (though each iteration requires more computation); hence they are suitable for small to medium scale problems requiring high accuracy solutions (e.g., SfM and photometric stereo in computer vision). Representatives of these algorithms include the damped Newton method [19], Wiberg(ℓ_2) [108], LM_S and LM_M of Chen [36] and LM_GN, which is a variant of LM_M using Gauss-Newton (GN) to approximate the Hessian function.

As these methods are of special importance in developing our PARSuMi algorithm, we conduct extensive numerical evaluations of these algorithms in Section 5.3 to understand their pros and cons as well as the key factors that lead to some of them finding global optimal solutions more often than others.

In addition, there are a few other works in each category that take into account the corruption problem by changing the quadratic penalty term of (5.8) into ℓ_1-norm or Huber function

$$\min_{U,V} \quad \left\| \mathcal{P}_\Omega(UV^T - \widehat{W}) \right\|_1 , \tag{5.10}$$

$$\min_{U,V} \quad \sum_{(ij)\in\Omega} \text{Huber}\big((UV^T - \widehat{W})_{ij}\big). \tag{5.11}$$

Notable algorithms to solve these formulations include alternating linear programming (ALP) and alternating quadratic programming (AQP) in Ke and Kanade [80], GRASTA

[1] Our experiment on synthetic data shows that the strong Wolfe line search adopted by Srebro and Jaakkola [127] and Wen and Yin [147] somewhat ameliorates the issue, though it does not seem to help much on real data.

[71] that extends GROUSE, as well as Wiberg ℓ_1 [58] that uses a second order Wiberg-like iteration. While it is well known that the ℓ_1-norm or Huber penalty term can better handle outliers, and the models (5.10) and (5.11) are seen to be effective in some problems, there is not much reason for a "convex" relaxation of the ℓ_0 pseudo-norm[1], since the rank constraint is already highly non-convex. Empirically, we find that ℓ_1-norm penalty offers poor denoising ability to dense noise and also suffers from "Robin Hood" attribute. Comparison with this class of methods will be given later in Section 5.5, which shows that our method can better handle noise and corruptions.

The practical advantage of ℓ_0 over ℓ_1 penalty is well illustrated in Xiong et al. [151], where Xiong et al. proposed an ℓ_0-based robust matrix factorization method which deals with corruptions and a given rank constraint. Our work is similar to Xiong et al. [151] in that we both eschew the convex surrogate ℓ_1-norm in favor of using the ℓ_0-norm directly. However, our approach treats both corruptions and missing data. More importantly, our treatment of the problem is different and it results in a convergence guarantee that covers the algorithm of Xiong et al. [151] as a special case; this will be further explained in Section 5.4.

5.2.3 Emerging theory for matrix factorization

As we mentioned earlier, a fundamental drawback of matrix factorization methods for low rank matrix completion is the lack of proper theoretical foundation. However, thanks to the better understanding of low-rank structures nowadays, some theoretical analysis of this problem slowly emerges. This class of methods are essentially designed for solving noisy matrix completion problem with an explicit rank constraint, i.e.,

$$\min_{W} \left\{ \frac{1}{2} \left\| \mathcal{P}_{\Omega}(W - \widehat{W}) \right\|^2 \,\middle|\, \operatorname{rank}(W) \leq r \right\}. \tag{5.12}$$

From a combinatorial-algebraic perspective, Kiraly and Tomioka [84] provided a sufficient and necessary condition on the existence of an unique rank-r solution to (5.12).

[1] The cardinality of non-zero entries, which strictly speaking is not a norm.

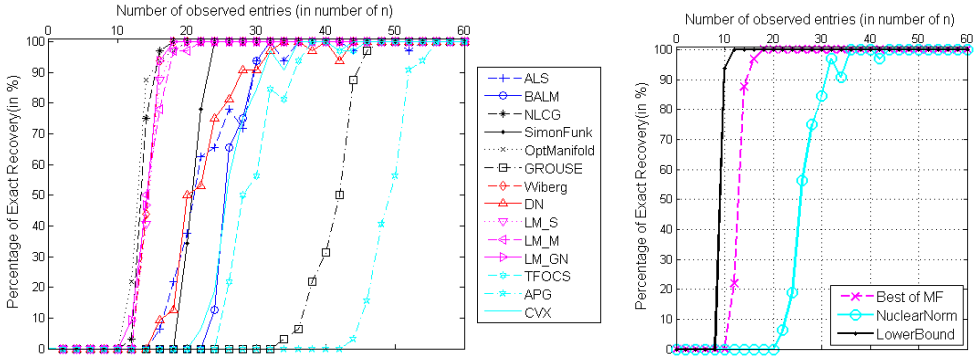


Figure 5.2: Exact recovery with **increasing number of random observations**. Algorithms (random initialization) are evaluated on 100 randomly generated rank-4 matrices of dimension 100×100. The number of observed entries increases from 0 to $50n$. To account for small numerical error, the result is considered "exact recovery" if the RMSE of the recovered entries is smaller than 10^{-3}. On the left, CVX [66], TFOCS [11] and APG [134] (in cyan) solves the nuclear norm based matrix completion (5.4), everything else aims to solve matrix factorization (5.8). On the right, the best solution of MF across all algorithms is compared to the CVX solver for nuclear norm minimization (solved with the highest numerical accuracy) and a lower bound (below the bound, the number of samples is smaller than r for at least a row or a column).

It turns out that if the low-rank matrix is *generic*, then the *unique completability* depends only on the support of the observations Ω. This suggests that the incoherence and random sampling assumptions typically required by various nuclear norm minimization methods may limit the portion of problems solvable by the latter to only a small subset of those solvable by matrix factorization methods.

Around the same time, Wang and Xu [142] studied the stability of matrix factorization under arbitrary noise. They obtained a stability bound for the optimal solution of (5.12) around the ground truth, which turns out to be better than the corresponding bound for nuclear norm minimization in Candes and Plan [21] by a scale of $\sqrt{\min(m,n)}$ (in Big-O sense). The study however bypassed the practical problem of how to obtain the global optimal solution for this non-convex problem.

This gap is partially closed by the recent work of Jain et al. [76], in which the global minimum of (5.12) can be obtained up to an accuracy ϵ with $O(\log 1/\epsilon)$ iterations using a slight variation of the ALS scheme. The guarantee requires the observation to be noiseless, sampled uniformly at random and the underlying subspace of W needs

to be incoherent—basically all assumptions in the convex approach—yet still requires slightly more observations than that for nuclear norm minimization. It does not however touch on when the algorithm is able to find the global optimal solution when the data is noisy. Despite not achieving stronger theoretical results nor under weaker assumptions than the convex relaxation approach, this is the first guarantee of its kind for matrix factorization. Given its more effective empirical performance, we believe that there is great room for improvement on the theoretical front. A secondary contribution of the results in this chapter is to find the potentially "right" algorithm or rather constituent elements of algorithm for theoreticians to look deeper into.

5.3 Numerical evaluation of matrix factorization methods

To better understand the performance of different methods, we compare the following attributes quantitatively for all three categories of approaches that solve (5.8) or (5.9)[1]:

Sample complexity Number of samples required for exact recovery of random uniformly sampled observations in random low-rank matrices, an index typically used to quantify the performance of nuclear norm based matrix completion.

Hits on global optimal[synthetic] The proportion of random initializations that lead to the global optimal solution on random low rank matrices with (a) increasing Gaussian noise, (b) exponentially decaying singular values.

Hits on global optimal[SfM] The proportion of random initializations that lead to the global optimal solution on the Oxford Dinosaur sequence [19] used in the SfM community.

The sample complexity experiment in Figure 5.2 shows that the best performing matrix factorization algorithm attains exact recovery with the number of observed entries at roughly 18%, while CVX for nuclear norm minimization needs roughly 36% (even worse for numerical solvers such as TFOCS and APG). This seems to imply that

[1] As a reference, we also included nuclear norm minimization that solve (5.4) where applicable.

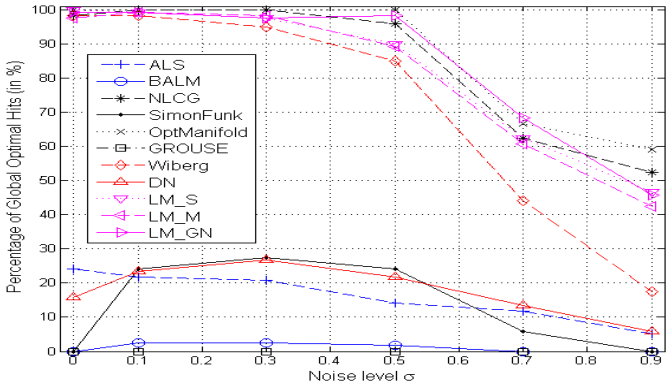


Figure 5.3: Percentage of hits on global optimal with **increasing level of noise**. 5 rank-4 matrices are generated by multiplying two standard Gaussian matrices of dimension 40×4 and 4×60. 30% of entries are uniformly picked as observations with additive Gaussian noise $N(0, \sigma)$. 24 different random initialization are tested for each matrix. The "global optimal" is assumed to be the solution with lowest objective value across all testing algorithm and all initializations.

the sample requirement for MF is fundamentally smaller than that of nuclear norm minimization. As MF assumes known rank of the underlying matrix while nuclear norm methods do not, the results we observe are quite reasonable. In addition, among different MF algorithms, some perform much better than others. The best few of them achieve something close to the lower bound[1]. This corroborates our intuition that MF is probably a better choice for problems with known rank.

From Figure 5.3 and 5.4, we observe that the following classes of algorithms, including LM_X series [36], Wiberg [108], Non-linear Conjugate Gradient method (NLCG) [127] and the curvilinear search on Stiefel manifold (OptManifold [147]) perform significantly better than others in reaching the global optimal solution despite their non-convexity. The percentage of global optimal hits from random initialization is promising even when the observations are highly noisy or when the condition number of the underlying matrix is very large[2].

[1]The lower bound is given by the percentage of randomly generated data that have at least one column or row having less than r samples. Clearly, having at least r samples for every column and row is a necessary condition for exact recovery.

[2]When $\alpha = 3.5$ in Figure 5.4, r^{th} singular value is almost as small as the spectral norm of the input noise.

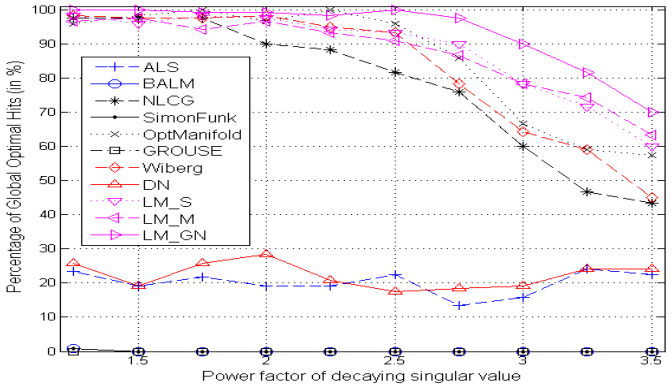


Figure 5.4: Percentage of hits on global optimal for **ill-conditioned low-rank matrices**. Data are generated in the same way as in Fig. 5.3 with $\sigma = 0.05$, except that we further take SVD and rescale the i^{th} singular value according to $1/\alpha^i$. The Frobenious norm is normalized to be the same as the original low-rank matrix. The exponent α is given on the horizontal axis.

The common attribute of the four algorithms is that they are all applied to the model (5.9) which parameterize the factor V as a function of U and then optimize over U alone. This parameterization essentially reduces the problem to finding the best subspace that fits the data. What is slightly different between them is the way they avoid local minima. OptManifold and NLCG adopt a Strong Wolfe line search that allows the algorithm to jump from one valley to another with long step sizes. The second order methods approximate each local neighborhood with a convex quadratic function and jump directly to the minimum of the approximation, thus rendering them liable to jump in an unpredictable fashion[1] until they reach a point in the basin of convergence where the quadratic approximation makes sense.

The difference in how the local minima are avoided appears to matter tremendously on the SfM experiment (see Figure 5.5). We observe that only the second order methods achieve global optimal solution frequently, whereas the Strong Wolfe line search adopted by both OptManifold and NLCG does not seem to help much on the real data experiment like it did in simulation with randomly generated data. Indeed, neither approach reaches the global optimal solution even once in the hundred runs, though they

[1] albeit always reducing the objective value due to the search on the Levenberg-Marquadt damping factor

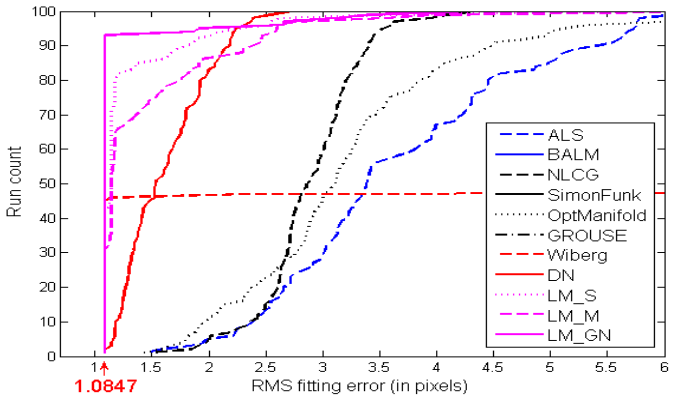
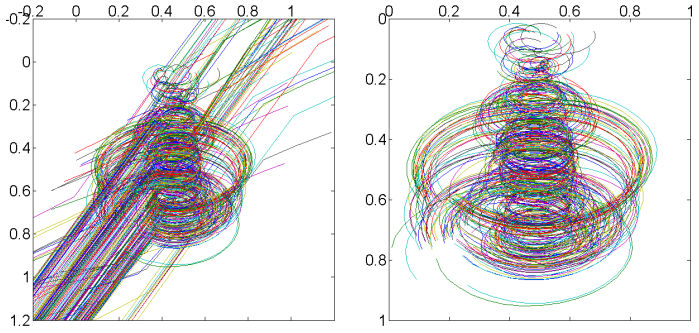


Figure 5.5: Accumulation histogram on the pixel RMSE for 100 randomly initialized runs are conducted for each algorithm on Dinosaur sequence. The curve summarizes how many runs of each algorithm corresponds to the global optimal solution (with pixel RMSE 1.0847) on the horizontal axis. Note that the input pixel coordinates are normalized to between $[0, 1]$ for experiments, but to be comparable with [19], the objective value is scaled back to the original size.

are rather close in quite a few runs. Despite these close runs, we remark that in applications like SfM, it is important to actually reach the global optimal solution. Due to the large amount of missing data in the matrix, even slight errors in the sampled entries can cause the recovered missing entries to go totally haywire with a seemingly good local minimum (see Figure 5.6). We thus refrain from giving any credit to local minima even if the $\mathrm{RMSE}_{\text{visible}}$ error (defined in (5.13)) is very close to that of the global minimum.

$$\mathrm{RMSE}_{\text{visible}} := \frac{\|\mathcal{P}_\Omega(W_{\text{recovered}} - \widehat{W})\|}{\sqrt{|\Omega|}}. \tag{5.13}$$

Another observation is that LM_GN seems to work substantially better than other second-order methods with subspace or manifold parameterization, reaching global minimum 93 times out of the 100 runs. Compared to LM_S and LM_M, the only difference is the use of Gauss-Newton approximation of the Hessian. According to the analysis in Chen [38], the Gauss-Newton Hessian provides the only non-negative convex quadratic approximation that preserves the so-called "zero-on-$(n-1)$-D" structure of a class of nonlinear least squares problems, for which (5.8) can be formulated. Compared to the Wiberg algorithm that also uses Gauss-Newton approximation, the advantage of



(a) Local minimum (b) Global minimum

Figure 5.6: Comparison of the feature trajectories corresponding to a local minimum and global minimum of (5.8), given partial uncorrupted observations. Note that $\text{RMSE}_{\text{visible}} = 1.1221$pixels in (a) and $\text{RMSE}_{\text{visible}} = 1.0847$pixels in (b). The latter is precisely the reported global minimum in Buchanan and Fitzgibbon [19], Okatani and Deguchi [108] and Chen [36]. Despite the tiny difference in $\text{RMSE}_{\text{visible}}$, the filled-in values for missing data in (a) are far off.

LM_GN is arguably the better global convergence due to the augmentation of the LM damping factor. Indeed, as we verify in the experiment, Wiberg algorithm fails to converge at all in most of its failure cases. The detailed comparisons of the second order methods and their running time on the Dinosaur sequence are summarized in Table 5.2. Part of the results replicate that in Chen [36]; however, Wiberg algorithm and LM_GN have not been explicitly compared previously. It is clear from the Table that LM_GN is not only better at reaching the optimal solution, but also computationally cheaper than other methods which require explicit computation of the Hessian[1].

To summarize the key findings of our experimental evaluation, we observe that: (a) the fixed-rank MF formulation requires less samples than nuclear norm minimization to achieve exact recovery; (b) the compact parameterization on the subspace, strong line search or second order update help MF algorithms in avoiding local minima in high noise, poorly conditioned matrix setting; (c) LM_GN with Gauss-Newton update is able to reach the global minimum with a very high success rate on a challenging real SfM data sequence.

[1] Wiberg takes longer time mainly because it sometimes does not converge and exhaust the maximum number of iterations.

	DN	Wiberg	LM_S	LM_M	LM_GN
No. of hits at global min.	2	46	42	32	93
No. of hits on stopping condition	75	47	99	93	98
Average run time(sec)	324	837	147	126	40
No. of variables	(m+n)r	(m-r)r	mr	(m-r)r	(m-r)r
Hessian	Yes	Gauss-Newton	Yes	Yes	Gauss-Newton
LM/Trust Region	Yes	No	Yes	Yes	Yes
Largest Linear system to solve	$[(m+n)r]^2$	$\|\Omega\| \times mr$	$mr \times mr$	$[(m-r)r]^2$	$[(m-r)r]^2$

Table 5.2: Comparison of various second order matrix factorization algorithms

We remark that while getting global optimal solution is important in applications like SfM, it is much less important in other applications such as collaborative filtering and feature learning etc. In those applications, the data set is bigger, but sparser and noisier and the low-rank model itself may be inaccurate in the first place. Getting a globally optimal solution may not correspond to a better estimate of unobserved data (aka smaller generalization error). Therefore, getting a somewhat reasonable solution really fast and making it online updatable are probably more important priorities. In this light, incremental algorithms like SimonFunk and GROUSE would be more appropriate, despite their inability to attain globally (perhaps even locally) optimal solution.

5.4 Proximal Alternating Robust Subspace Minimization for (5.3)

Our proposed PARSuMi method for problem (5.3) works in two stages. It first obtains a good initialization from an efficient convex relaxation of (5.3), which will be described in Section 5.4.5. This is followed by the minimization of the low rank matrix W and the sparse matrix E alternatingly until convergence. The efficiency of our PARSuMi method depends on the fact that the two inner minimizations of W and E admit efficient solutions, which will be derived in Sections 5.4.1 and 5.4.2 respectively. Specifically,

in step k, we compute W^{k+1} from

$$\begin{aligned} &\min_{W} \frac{1}{2}\|H \circ (W - \widehat{W} + E^k)\|^2 + \frac{\beta_1}{2}\|H \circ (W - W^k)\|^2 \\ &\text{subject to} \quad \operatorname{rank}(W) \leq r, \end{aligned} \tag{5.14}$$

and E^{k+1} from

$$\begin{aligned} &\min_{E} \frac{1}{2}\|H \circ (W^{k+1} - \widehat{W} + E)\|^2 + \frac{\beta_2}{2}\|E - E^k\|^2 \\ &\text{subject to} \quad \|E\|_0 \leq N_0,\ \|E\| \leq K_E,\ E \in \mathbb{R}_{\Omega}^{m \times n}, \end{aligned} \tag{5.15}$$

where H is defined as in (5.2). Note that the above iteration is different from applying a direct alternating minimization of (5.3). We have added the proximal regularization terms $\|H \circ (W - W^k)\|^2$ and $\|E - E^k\|^2$ to make the objective functions in the subproblems coercive and hence ensuring that W^{k+1} and E^{k+1} are well defined. As is shown in Attouch et al. [3], the proximal terms are critical to ensure the critical point convergence of the sequence. We provide the formal critical point convergence proof of our algorithm in Section 5.4.4.

5.4.1 Computation of W^{k+1} in (5.14)

Our solution for (5.14) consists of two steps. We first transform the rank-constrained minimization (5.14) into an equivalent (which we will show later) subspace fitting problem, then solve the new formulation using LM_GN.

Motivated by the findings in Section 5.3 where the most successful algorithms for solving (5.12) are based on the formulation (5.9), we will now derive a similar equivalent reformulation of (5.14). Our reformulation of (5.14) is motivated by the N-parametrization of (5.12) due to Chen [36], who considered the task of matrix completion as finding the best subspace to fit the partially observed data. In particular, Chen

proposes to solve (5.12) using

$$\min_{N} \left\{ \frac{1}{2} \sum_{i} \hat{w}_i^T (I - \mathbb{P}_i) \hat{w}_i \middle| N^T N = I \right\} \tag{5.16}$$

where N is a $m \times r$ matrix whose column space is the underlying subspace to be reconstructed, N_i is N but with those rows corresponding to the missing entries in column i removed. $\mathbb{P}_i = N_i N_i^+$ is the projection onto $\mathrm{span}(N_i)$ with N_i^+ being the Moore-Penrose pseudo inverse of N_i, and the objective function minimizes the sum of squares distance between $\hat{w}_i$ to $\mathrm{span}(N_i)$, where $\hat{w}_i$ is the vector of observed entries in the i^{th} column of $\widehat{W}$.

5.4.1.1 N-parameterization of the subproblem (5.14)

First define the matrix $\overline{H} \in \mathbb{R}^{m \times n}$ as follows:

$$\overline{H}_{ij} = \begin{cases} \sqrt{1 + \beta_1} & \text{if } (i,j) \in \Omega \\ \sqrt{\lambda + \lambda\beta_1} & \text{if } (i,j) \notin \Omega. \end{cases} \tag{5.17}$$

Let $B^k \in \mathbb{R}^{m \times n}$ be the matrix defined by

$$B_{ij} = \begin{cases} \frac{1}{\sqrt{1+\beta_1}} (\widehat{W}_{ij} - E_{ij}^k + \beta_1 W_{ij}^k) & \text{if } (i,j) \in \Omega \\ \frac{\lambda\beta_1}{\sqrt{\lambda+\lambda\beta_1}} W_{ij}^k & \text{if } (i,j) \notin \Omega. \end{cases} \tag{5.18}$$

Define the diagonal matrices $\mathbb{D}_i \in \mathbb{R}^{m \times m}$ to be

$$\mathbb{D}_i = \mathrm{diag}(\overline{H}_i), \quad i = 1, \ldots, n \tag{5.19}$$

where $\overline{H}_i$ is the ith column of $\overline{H}$. It turns out that the N-parameterization for the regularized problem (5.14) has a similar form as (5.16), as shown below.

Proposition 5.1 (Equivalence of subspace parameterization)**.** *Let* $\mathbb{Q}_i(N) = \mathbb{D}_i N (N^T \mathbb{D}_i^2 N)^{-1} N^T \mathbb{D}_i$*, which is the* $m \times m$ *projection matrix onto the column space of* $\mathbb{D}_i N$*. The problem* (5.14)

is equivalent to the following problem:

$$\begin{aligned} \min_{N} \quad & f(N) := \frac{1}{2}\sum_{i=1}^{n} \|B_i^k - \mathbb{Q}_i(N)B_i^k\|^2 \\ \text{subject to} \quad & N^T N = I,\ N \in \mathbb{R}^{m\times r} \end{aligned} \tag{5.20}$$

where B_i^k is the ith columns of B^k. If N_* is an optimal solution of (5.20)*, then W^{k+1}, whose columns are defined by*

$$W_i^{k+1} = \mathbb{D}_i^{-1}\mathbb{Q}_i(N_*)B_i^k, \tag{5.21}$$

is an optimal solution of (5.14)*.*

Proof. We can show by some algebraic manipulations that the objective function in (5.14) is equal to

$$\frac{1}{2}\|\overline{H} \circ W - B^k\|^2 + \text{constant}$$

Now note that we have

$$\{W \in \mathbb{R}^{m\times n} \mid \text{rank}(W) \le r\} = \{NC \mid N \in \mathbb{R}^{m\times r}, C \in \mathbb{R}^{r\times n}, N^T N = I\}.$$

Thus the problem (5.14) is equivalent to

$$\min_{N}\{f(N) \mid N^T N = I, N \in \mathbb{R}^{m\times r}\} \tag{5.22}$$

where

$$f(N) := \min_{C} \frac{1}{2}\|\overline{H} \circ (NC) - B^k\|^2.$$

To derive (5.20) from the above, we need to obtain $f(N)$ explicitly as a function of N. For a given N, the unconstrained minimization problem over C in $f(N)$ has a strictly convex objective function in C, and hence the unique global minimizer satisfies the

following optimality condition:

$$N^T((\overline{H} \circ \overline{H}) \circ (NC)) = N^T(\overline{H} \circ B^k). \tag{5.23}$$

By considering the ith column C_i of C, we get

$$N^T \mathbb{D}_i^2 N C_i = N^T \mathbb{D}_i B_i^k, \quad i = 1, \ldots, n. \tag{5.24}$$

Since N has full column rank and D^i is positive definite, the coefficient matrix in the above equation is nonsingular, and hence

$$C_i = (N^T \mathbb{D}_i^2 N)^{-1} N^T \mathbb{D}_i B_i^k.$$

Now with the optimal C_i above for the given N, we can show after some algebra manipulations that $f(N)$ is given as in (5.20). □

We can see that when $\beta_1 \downarrow 0$ in (5.20), then the problem reduces to (5.16), with the latter's $\hat{w}_i$ appropriately modified to take into account of E^k. Also, from the above proof, we see that the N-parameterization reduces the feasible region of W by restricting W to only those potential optimal solutions among the set of W satisfying the expression in (5.21). This seems to imply that it is not only equivalent but also advantageous to optimize over N instead of W. While we have no theoretical justification of this conjecture, it is consistent with our experiments in Section 5.3 which show the superior performance of those algorithms using subspace parameterization in finding global minima and vindicates the design motivations of the series of LM_X algorithms in Chen [36].

5.4.1.2 LM_GN updates

Now that we have shown how to handle the regularization term and validated the equivalence of the transformation, the steps to solve (5.14) essentially generalize those of LM_GN (available in Section 3.2 and Appendix A of Chen [38]) to account for the gen-

eral mask H. The derivations of the key formulae and their meanings are given in this section.

In general, Levenberg-Marquadt solves the non-linear problem with the following sum-of-squares objective function

$$\mathcal{L}(x) = \frac{1}{2} \sum_{i=1:n} \|y_i - f_i(x)\|^2, \tag{5.25}$$

by iteratively updating x as follows:

$$x \leftarrow x + (J^T J + \lambda I)^{-1} J^T \mathbf{r},$$

where $J = [J_1; \ldots; J_n]$ is the Jacobian matrix where J_i is the Jacobian matrix of f_i; $\mathbf{r}$ is the concatenated vector of residual $r_i := y_i - f_i(x)$ for all i, and λ is the damping factor that interpolates between Gauss-Newton update and gradient descent. We may also interpret the iteration as a Damped Newton method with a first order approximation of the Hessian matrix using $H \approx J^T J$.

Note that the objective function of (5.20) can be expressed in the form of (5.25) by taking $x := \mathrm{vec}(N)$, data $y_i := B_i^k$, and function

$$f_i(x := \mathrm{vec}(N)) = \mathbb{Q}_i(N) B_i^k = \mathbb{Q}_i y_i$$

Proposition 5.2. *Let $\mathcal{T} \in \mathbb{R}^{mr \times mr}$ be the permutation matrix such that $\mathrm{vec}(X^T) = \mathcal{T}\mathrm{vec}(X)$ for any $X \in \mathbb{R}^{m \times r}$. The Jacobian of $f_i(x) = \mathbb{Q}_i(N) y_i$ is given as follows:*

$$J_i(x) = (\mathbb{A}_i^T y_i)^T \otimes ((I - \mathbb{Q}_i)\mathbb{D}_i) + \ [(\mathbb{D}_i r_i)^T \otimes \mathbb{A}_i]\mathcal{T}. \tag{5.26}$$

Also $J^T J = \sum_{i=1}^n J_i^T J_i$, $J^T r = \sum_{i=1}^n J_i^T r_i$, where

$$J_i^T J_i = (\mathbb{A}_i^T y_i y_i^T \mathbb{A}_i) \otimes (\mathbb{D}_i (I - \mathbb{Q}_i)\mathbb{D}_i) + \mathcal{T}^T [(\mathbb{D}_i r_i r_i^T \mathbb{D}_i) \otimes (\mathbb{A}_i^T \mathbb{A}_i)]\mathcal{T} \tag{5.27}$$

$$J_i^T r_i = \mathrm{vec}(\mathbb{D}_i r_i (\mathbb{A}_i^T y_i)^T). \tag{5.28}$$

In the above, $\otimes$ denotes the Kronecker product.

Proof. Let $\mathbb{A}_i = \mathbb{D}_i N (N^T \mathbb{D}_i^2 N)^{-1}$. Given sufficiently small δN, we can show that the directional derivative of f_i at N along δN is given by

$$f_i'(N + \delta N) \quad = \quad (I - \mathbb{Q}_i)\mathbb{D}_i \delta N \mathbb{A}_i^T y_i + \mathbb{A}_i \delta N^T \mathbb{D}_i r_i.$$

By using the property that $\mathrm{vec}(AXB) = (B^T \otimes A)\mathrm{vec}(X)$, we have

$$\begin{aligned} \mathrm{vec}(f_i'(N + \delta N)) \quad &= \quad [(\mathbb{A}_i^T y_i)^T \otimes ((I - \mathbb{Q}_i)\mathbb{D}_i)]\mathrm{vec}(\delta N) \\ &\quad + [(\mathbb{D}_i r_i)^T \otimes \mathbb{A}_i]\mathrm{vec}(\delta N^T) \end{aligned}$$

From here, the required result in (5.26) follows.

To prove (5.27), we make use the following properties of Kronecker product: $(A \otimes B)(C \otimes D) = (AC) \otimes (BD)$ and $(A \otimes B)^T = A^T \otimes B^T$. By using these properties, we see that $J_i^T J_i$ has 4 terms, with two of the terms contain the Kronecker products involving $\mathbb{D}_i(I - \mathbb{Q}_i)\mathbb{A}_i$ or its transpose. But we can verify that $\mathbb{Q}_i \mathbb{A}_i = \mathbb{A}_i$ and hence those two terms become 0. The remaining two terms are those appearing in (5.27) after using the fact that $(I - \mathbb{Q}_i)^2 = I - \mathbb{Q}_i$. Next we prove (5.28). We have

$$J_i^T r_i = \mathrm{vec}(\mathbb{D}_i(I - \mathbb{Q}_i) r_i (\mathbb{A}_i^T y_i)^T) + \mathfrak{T}^T \mathrm{vec}(\mathbb{A}_i^T r_i r_i^T \mathbb{D}_i).$$

By noting that $\mathbb{A}_i^T r_i = 0$ and $\mathbb{Q}_i r_i = 0$, we get the required result in (5.28). □

The complete procedure of solving (5.14) is summarized in Algorithm 1.

5.4.2 Sparse corruption recovery step (5.15)

In the sparse corruption step, we need to solve the ℓ_0-constrained least squares minimization (5.15). This problem is combinatorial in nature, but fortunately, for our problem, we show that a closed-form solution can be obtained. Let $x := \mathcal{P}_\Omega(E)$. Observe

Algorithm 1 Leverberg-Marquadt method for (5.14)

Input: $\widehat{W}, E^k, W^k, \Omega$, objective function $\mathcal{L}(x)$ and initial N^k; numerical parameter λ, $\rho > 1$.
Initialization: Compute $y_i = B_i^k$ for $i = 1, ..., n$, and $x^0 = \text{vec}(N^k)$, $j = 0$.
while not converged **do**
1. Compute $J^T\mathbf{r}$ and J^TJ using (5.28) and(5.27).
2. Compute $\Delta x = (J^TJ + \lambda I)^{-1}J^Tr$
while $\mathcal{L}(x + \Delta x) < \mathcal{L}(x)$ **do**
(1) $\lambda = \rho\lambda$.
(2) $\Delta x = (J^TJ + \lambda I)^{-1}J^Tr$.
end while
3. $\lambda = \lambda/\rho$.
4. Orthogonalize $N = \text{orth}[\text{reshape}(x^j + \Delta x)]$.
5. Update $x^{j+1} = \text{vec}(N)$.
6. Iterate $j = j + 1$
end while
Output: $N^{k+1} = N$, W^{k+1} using (5.21) with N^{k+1} replacing N_*.

that (5.15) can be expressed in the following equivalent form:

$$\min_x \Big\{ \|x - b\|^2 \mid \|x\|_0 \le N_0,\ \|x\|^2 - K_E^2 \le 0 \Big\} \tag{5.29}$$

where $b = \mathcal{P}_\Omega(\widehat{W} - W^{k+1} + \beta_2 E^k)/(1 + \beta_2)$.

Proposition 5.3. *Let I be the set of indices of the N_0 largest (in magnitude) component of b. Then the nonzero components of the optimal solution x of* (5.29) *is given by*

$$x_I = \begin{cases} K_E b_I/\|b_I\| & \text{if } \|b_I\| > K_E \\ b_I & \text{if } \|b_I\| \le K_E. \end{cases} \tag{5.30}$$

Proof. Given a subset I of $\{1, \ldots, |\Omega|\}$ with cardinality at most N_0 such that $b_I \neq 0$. Let $J = \{1, \ldots, |\Omega|\}\backslash I$. Consider the problem (5.29) for $x \in \mathbb{R}^{|\Omega|}$ supported on I, we get the following:

$$v_I := \min_{x_I} \Big\{ \|x_I - b_I\|^2 + \|b_J\|^2 \mid \|x_I\|^2 - K_E^2 \le 0 \Big\},$$

Algorithm 2 Closed-form solution of (5.15)

Input: $\widehat{W}, W^{k+1}, E^k, \Omega$.
1. Compute b using (5.29).
2. Compute x using (5.30).

Output: $E^{k+1} = P^*_\Omega(x)$.

which is a convex minimization problem whose optimality conditions are given by

$$x_I - b_I + \mu\, x_I = 0,\ \mu(\|x_I\|^2 - K_E^2) = 0,\ \mu \geq 0$$

where μ is the Lagrange multiplier for the inequality constraint. First consider the case where $\mu > 0$. Then we get $x_I = K_E b_I/\|b_I\|$, and $1 + \mu = \|b_I\|/K_E$ (hence $\|b_I\| > K_E$). This implies that $v_I = \|b\|^2 + K_E^2 - 2\|b_I\|K_E$. On the other hand, if $\mu = 0$, then we have $x_I = b_I$ and $v_I = \|b_J\|^2 = \|b\|^2 - \|b_I\|^2$. Hence

$$v_I = \begin{cases} \|b\|^2 + K_E^2 - 2\|b_I\|K_E & \text{if } \|b_I\| > K_E \\ \|b\|^2 - \|b_I\|^2 & \text{if } \|b_I\| \leq K_E. \end{cases}$$

In both cases, it is clear that v_I is minimized if $\|b_I\|$ is maximized. Obviously $\|b_I\|$ is maximized if I is chosen to be the set of indices corresponding to the N_0 largest components of b. □

The procedure to obtain the optimal solution of (5.15) is summarized in Algorithm 2. We remark that this is a very special case of ℓ_0-constrained optimization; the availability of the exact closed form solution depends on both terms in (5.15) being decomposable into individual (i, j) term. In general, if we change the operator $M \to H \circ M$ in (5.15) to a general linear transformation (e.g., a sensing matrix in compressive sensing), or change the norm $\|\cdot\|$ of the proximal term to some other norm such as spectral norm or nuclear norm, then the problem becomes NP-hard.

Algorithm 3 Proximal Alternating Robust Subspace Minimization (PARSuMi)

Input: Observed data $\widehat{W}$, sample mask Ω, parameter r, N_0. Initialization W^0 and E^0 (typically by Algorithm 5 described in Section 5.4.5), $k = 0$.
repeat
1. Solve (5.14) using Algorithm 1 with W^k,E^k,N^k, obtain updates W^{k+1} and N^{k+1}
2. Solve (5.15) using Algorithm 2 with W^{k+1},E^k obtain updates E^{k+1}.

until $\|W^{k+1} - W^k\| < \|W^k\| \cdot 10^{-6}$ and $\|E^{k+1} - E^k\| < \|E^k\| \cdot 10^{-6}$
Output: Accumulation points $\overline{W}$ and $\overline{E}$

5.4.3 Algorithm

Our Proximal Alternating Robust Subspace Minimization method is summarized in Algorithm 3. Note that we do not need to know the exact cardinality of the corrupted entries; N_0 can be taken as an upper bound of allowable number of corruption. As a rule of thumb, 10%-15% of $|\Omega|$ is a reasonable size. The surplus in N_0 will only label a few noisy samples as corruptions, which should not affect the recovery of either W or E, so long as the remaining $|\Omega| - N_0$ samples are still sufficient.

5.4.4 Convergence to a critical point

In this section, we show the convergence of Algorithm 3 to a critical point. This critical point guarantee is of theoretical significance because as far as we know, our critical point guarantee produces a stronger result compared to the widely used alternating minimization or block coordinate descent (BCD) methods in computer vision problems. A relevant and interesting comparison is the Bilinear Alternating Minimization(BALM) [46] work, where the critical point convergence of the alternating minimization is proven in Xavier et al. [150]. The proof is contingent on the smoothness of the Stiefel manifold. In contrast, our proposed proximal alternating minimization framework based on Attouch et al. [3] is more general in the sense that convergence to a critical point can be established for non-smooth and non-convex objective functions or constraints.

We start our convergence proof by first defining an equivalent formulation of (5.3) in terms of closed, bounded sets. The convergence proof is then based on the indicator

functions for these closed and bounded sets, which have the key lower semicontinuous property.

Let $K_W = 2\|\widehat{W}\| + K_E$. Define the closed and bounded sets:

$$\begin{aligned} \mathcal{W} &= \{W \in \mathbb{R}^{m\times n} \mid \mathrm{rank}(W) \le r, \|H \circ W\| \le K_W\} \\ \mathcal{E} &= \{E \in \mathbb{R}^{m\times n}_\Omega \mid \|E\|_0 \le N_0, \|E\| \le K_E\}. \end{aligned}$$

We will first show that (5.3) is equivalent to the problem given in the next proposition.

Proposition 5.4. *The problem* (5.3) *is equivalent to the following problem:*

$$\begin{aligned} \min \quad & f(W,E) := \tfrac{1}{2}\|H \circ (W + E - \widehat{W})\|^2 \\ \textit{s.t.} \quad & W \in \mathcal{W}, E \in \mathcal{E}. \end{aligned} \tag{5.31}$$

Proof. Observe the only difference between (5.3) and (5.31) is the inclusion of the bound constraint on $\|H \circ W\|$ in (5.31). To show the equivalence, we only need to show that any minimizer (W^*, E^*) of (5.3) must satisfy the bound constraint in $\mathcal{W}$. By definition, we know that

$$f(W^*, E^*) \le f(0,0) = \frac{1}{2}\|\widehat{W}\|^2.$$

Now for any (W, E) such that $\mathrm{rank}(W) \le r$, $E \in \mathcal{E}$ and $\|H \circ W\| > K_W$, we must have

$$\begin{aligned} \|H \circ (W + E - \widehat{W})\| &\ge \|H \circ W\| - \|H \circ (E - \widehat{W})\| \\ &> K_W - \|E\| - \|\widehat{W}\| \ge \|\widehat{W}\|. \end{aligned}$$

Hence $f(W,E) > \frac{1}{2}\|\widehat{W}\|^2 = f(0,0)$. This implies that we must have $\|H \circ W^*\| \le K_W$. □

Let $\mathcal{X}$ and $\mathcal{Y}$ be the finite-dimensional inner product spaces, $\mathbb{R}^{m\times n}$ and $\mathbb{R}^{m\times n}_\Omega$, re-

spectively. If we define $f : \mathcal{X} \to \mathbb{R} \cup \{\infty\}$, $g : \mathcal{Y} \to \mathbb{R} \cup \{\infty\}$ to be the following indicator functions,

$$f(x) = \delta_{\mathcal{W}}(x) = \begin{cases} 0 & \text{if } x \in \mathcal{W} \\ \infty & \text{otherwise} \end{cases}$$

$$g(y) = \delta_{\mathcal{E}}(y) = \begin{cases} 0 & \text{if } y \in \mathcal{E} \\ \infty & \text{otherwise} \end{cases}$$

then we can rewrite (5.31) as the following equivalent problem:

$$\underset{x,y}{\text{minimize}}\{L(x,y) := f(x) + g(y) + q(x,y)\} \tag{5.32}$$

where

$$q(x,y) = \frac{1}{2}\|Ax + By - c\|^2$$

and $A : \mathcal{X} \to \mathcal{X}$, $B : \mathcal{Y} \to \mathcal{X}$ are given linear maps defined by $A(x) = H \circ x$, $B(y) = H \circ y$, and $c = \widehat{W}$. Note that in this case, f are g are lower semicontinuous since indicator functions of closed sets are lower semicontinuous [117].

Consider the proximal alternating minimization outlined in Algorithm 4, as proposed in Attouch et al. [3]. The algorithm alternates between minimizing x and y, but with the important addition of the quadratic Moreau-Yoshida regularization term (which is also known as the proximal term) in each step. The importance of Moreau-Yoshida regularization for convex matrix optimization problems has been demonstrated and studied in Bin et al. [16], Liu et al. [99], Yang et al. [152]. For our non-convex, non-smooth setting here, the importance of the proximal term will become clear when we prove the convergence of Algorithm 4. The positive linear maps S and T in Algorithm 4 correspond to $(H \circ H)\circ$ and the identity map respectively. Note that our formulation is slightly more general than that of Attouch et al. [3] in which the positive linear maps S and T are simply the identity maps.

Algorithm 4 Proximal alternating minimization

Input: $(x_0, y_0) \in \mathcal{X} \times \mathcal{Y}$

repeat

1. $x^{k+1} = \arg\min\{L(x, y^k) + \frac{\beta_1}{2}\|x - x^k\|_S^2\}$
2. $y^{k+1} = \arg\min\{L(x^{k+1}, y) + \frac{\beta_2}{2}\|y - y^k\|_T^2\}$

until convergence

Output: Accumulation points $\overline{x}$ and $\overline{y}$

In the above, S and T are given positive definite linear maps, and $\|x - x^k\|_S^2 = \langle x - x^k, S(x - x^k)\rangle$, $\|y - y^k\|_T^2 = \langle y - y^k, T(y - y^k)\rangle$.

In Attouch et al. [3], the focus is on non-smooth and non-convex problems where $q(x, y)$ is a smooth function with Lipschitz continuous gradient on the domain $\{(x, y) \mid f(x) < \infty, g(y) < \infty\}$, and f and g are lower semicontinuous functions (not necessarily indicator functions) such that $L(x, y)$ satisfy a key property (known as the Kurdyka-Lojasiewicz (KL) property) at some limit point of $\{(x^k, y^k)\}$. Typically the KL property can be established for semi-algebraic functions based on abstract mathematical arguments. Once the KL property is established, convergence to a critical point is guaranteed by virtue of Theorem 9 in Attouch et al. [3][1]. The KL property also allows stronger property to be derived. For example, Theorem 11 gives the rate of convergence, albeit depending on some constants which are usually not known explicitly.

For our more specialized problem (5.31), the KL property can also be established, although the derivation is non-trivial. Here we prefer to present a less abstract and simpler convergence proof. For the benefit of those readers who do not wish to deal with abstract concepts, Theorem 5.1 is self-contained and does not require the understanding of the abstract KL property. Our result is analogous to that in Section 3.1 in Attouch et al. [3] which proved a weaker form of convergence to a critical point without invoking the KL property. But note that our proposed algorithm 4 involves the more general positive linear maps ($\|.\|_S$ and $\|.\|_T$) in the proximal regularization. We therefore provide Theorem 5.1 for this more general form of proximal regularization.

There are four parts to Theorem 5.1. Part(a) establishes the non-increasing mono-

[1] Thus, the critical point convergence for BALM follows automatically by identifying the Stiefel manifold as a semialgebraic object and therefore satisfying the KL property.

tonicity of the proximal regularized update. Leveraging on part(a), part(b) ensures the existence of the limits. Using Part(a), (b) and (c), (d) then shows the critical point convergence proof.

Theorem 5.1. *Let $\{(x^k, y^k)\}$ be the sequence generated by Algorithm 4. Then the following statements hold.*

(a) For all $k \geq 0$,

$$\begin{aligned} &\frac{1}{2}\|x_{k+1} - x_k\|_S^2 + \frac{1}{2}\|y_{k+1} - y_k\|_T^2 \\ &\leq L(x_k, y_k) - L(x_{k+1}, y_{k+1}) \end{aligned} \tag{5.33}$$

(b) $\sum_{k=0}^{\infty} \frac{1}{2}\|x_{k+1} - x_k\|_S^2 + \frac{1}{2}\|y_{k+1} - y_k\|_T^2 < \infty$. Hence $\lim_{k\to\infty} \|x_{k+1} - x_k\| = 0 = \lim_{k\to\infty} \|y_{k+1} - y_k\|$.

(c) Let $\Delta x_{k+1} = A^*B(y^{k+1} - y^k) - S(x_{k+1} - x_k)$ and $\Delta y_{k+1} = -T(y_{k+1} - y_k)$. Then

$$(\Delta x_{k+1}, \Delta y_{k+1}) \in \partial L(x_{k+1}, y_{k+1}) \tag{5.34}$$

where $\partial L(x, y)$ denotes the subdifferential of L at (x, y).

(d) The sequence $\{(x^k, y^k)\}$ has a limit point. Any limit point $(\bar{x}, \bar{y})$ is a stationary point of the problem (5.31)*. Moreover, $\lim_{k\to\infty} L(x_k, y_k) = L(\bar{x}, \bar{y}) = \inf_k L(x_k, y_k)$.*

Proof. (a) By the minimal property of x_{k+1}, we have

$$\begin{aligned} & L(x_{k+1}, y_k) + \frac{1}{2}\|x_{k+1} - x_k\|_S^2 \\ = \;& \Big(f(x_{k+1}) + q(x_{k+1}, y_k) + \frac{1}{2}\|x_{k+1} - x_k\|_S^2\Big) + g(y_k) \\ \leq \;& \Big(f(\xi) + q(\xi, y_k) + \frac{1}{2}\|\xi - x_k\|_G^2\Big) + g(y_k) \\ = \;& L(\xi, y_k) + \frac{1}{2}\|\xi - x_k\|_S^2 \quad \forall\, \xi \in \mathcal{X}. \end{aligned} \tag{5.35}$$

Similarly, by the minimal property of y_{k+1}, we have

$$L(x_{k+1}, y_{k+1}) + \frac{1}{2}\|y_{k+1} - y_k\|_T^2 \;\leq\; L(x_{k+1}, \eta) + \frac{1}{2}\|\eta - y_k\|_T^2 \quad \forall\, \eta \in \mathcal{Y} \tag{5.36}$$

By taking $\xi = x_k$ in (5.35) and $\eta = y_k$ in (5.36), we get the required result.

(b) We omit the proof since the results are easy consequences of the result in (a). Note that to establish $\lim_{k\to\infty} \|x_{k+1} - x_k\| = 0$, we used the fact that $\|x_{k+1} - x_k\|_S \to 0$ as $k \to \infty$, and that S is a positive definite linear operator. Similar remark also applies to $\{y^{k+1} - y^k\}$.

(c) The result in (5.34) follows from the minimal properties of x_{k+1} and y_{k+1} in Step 1 and 2 of Algorithm 4, respectively.

(d) Because $\|H \circ x^k\| \leq K_W$ and $\|y^k\| \leq K_E$, the sequence $\{(x^k, y^k)\}$ is bounded and hence it has a limit point. Let $(x_{k'}, y_{k'})$ be a convergent subsequence with limit $(\bar{x}, \bar{y})$. From (5.35), we have $\forall\, \xi \in \mathcal{X}$

$$\limsup_{k'\to\infty} f(x_{k'}) + q(\bar{x}, \bar{y}) \;\leq\; f(\xi) + q(\xi, \bar{y}) + \frac{1}{2}\|\xi - \bar{y}\|_S^2.$$

By taking $\xi = \bar{x}$, we get $\limsup_{k'\to\infty} f(x_{k'}) \leq f(\bar{x})$. Also, we have $\liminf_{k'\to\infty} f(x_{k'}) \geq f(\bar{x})$ since f is lower semicontinuous. Thus $\lim_{k'\to\infty} f(x_{k'}) = f(\bar{x})$. Similarly, by using (5.36), we can show that $\lim_{k'\to\infty} g(y_{k'}) = g(\bar{y})$. As a result, we have

$$\lim_{k'\to\infty} L(x_{k'}, y_{k'}) \;=\; L(\bar{x}, \bar{y}).$$

Since $\{L(x_k, y_k)\}$ is a nonincreasing sequence, the above implies that $\lim_{k\to\infty} L(x_k, y_k) = L(\bar{x}, \bar{y}) = \inf_k L(x_k, y_k)$. Now from (c), we have

$$(\Delta x_{k'}, \Delta y_{k'}) \;\in\; \partial L(x_{k'}, y_{k'}), \quad (\Delta x_{k'}, \Delta y_{k'}) \;\to\; (0, 0).$$

By the closedness property of ∂L [41, Proposition 2.1.5], we get $0 \in \partial L(\bar{x}, \bar{y})$. Hence $(\bar{x}, \bar{y})$ is a stationary point of L. □

5.4.5 Convex relaxation of (5.3) as initialization

Due to the non-convexity of the rank and ℓ_0 cardinality constraints, it is expected that the outcome of Algorithm 3 depends on initializations. A natural choice for the initialization of PARSuMi is the convex relaxation of both the rank and ℓ_0 function:

$$\min \left\{ f(W,E) + \lambda\|W\|_* + \gamma\|E\|_1 \mid W \in \mathbb{R}^{m\times n}, E \in \mathbb{R}^{m\times n}_\Omega \right\} \tag{5.37}$$

where $f(W,E) = \frac{1}{2}\|H \circ (W + E - \widehat{W})\|^2$, $\|\cdot\|_*$ is the nuclear norm, and λ and γ are regularization parameters.

Problem (5.37) can be solved efficiently by the quadratic majorization-APG (accelerated proximal gradient) framework proposed by Toh and Yun [134]. At the kth iteration with iterate $(\bar{W}^k, \bar{E}^k)$, the majorization step replaces (5.37) with a quadratic majorization of $f(W,E)$, so that W and E can be optimized independently, as we shall see shortly. Let $G^k = (H \circ H) \circ (\bar{W}^k + \bar{E}^k + \widehat{W})$. By some simple algebra, we have

$$\begin{aligned} f(W,E) - f(\bar{W}^k, \bar{E}^k) &= \frac{1}{2}\|H \circ (W - \bar{W}^k + E - \bar{E}^k)\|^2 \\ &\quad + \langle W - \bar{W}^k + E - \bar{E}^k, G^k \rangle \\ &\leq \|W - \bar{W}^k\|^2 + \|E - \bar{E}^k\|^2 + \langle W - \bar{W}^k + E - \bar{E}^k, G^k \rangle \\ &= \|W - \widetilde{W}^k\|^2 + \|E - \widetilde{E}^k\|^2 + \text{constant} \end{aligned}$$

where $\widetilde{W}^k = \bar{W}^k - G^k/2$ and $\widetilde{E}^k = \bar{E}^k - G^k/2$. At each step of the APG method, one minimizes (5.37) with $f(W,E)$ replaced by the above quadratic majorization. As the resulting problem is separable in W and E, we can minimize them separately, thus yielding the following two optimization problems:

$$W^{k+1} = \text{argmin}\ \frac{1}{2}\|W - \widetilde{W}^k\|^2 + \frac{\lambda}{2}\|W\|_* \tag{5.38}$$

$$E^{k+1} = \text{argmin}\ \frac{1}{2}\|E - \widetilde{E}^k\|^2 + \frac{\gamma}{2}\|E\|_1 \tag{5.39}$$

The main reason for performing the above majorization is because the solutions to

(5.38) and (5.39) can readily be found with closed-form solutions. For (5.38), the minimizer is given by the Singular Value Thresholding (SVT) operator. For (5.39), the minimizer is given by the well-known soft thresholding operator [47]. The APG algorithm, which is adapted from Beck and Teboulle [9] and analogous to that in Toh and Yun [134], is summarized below.

Algorithm 5 An APG algorithm for (5.37)

Input: Initialize $W^0 = \bar{W}^0 = 0$, $E^0 = \bar{E}^0 = 0$, $t_0 = 1$, $k = 0$
repeat
1. Compute $G^k = (H \circ H) \circ (\bar{W}^k + \bar{E}^k + \widehat{W})$, $\widetilde{W}^k$, $\widetilde{E}^k$.
2. Update W^{k+1} by applying the SVT on $\widetilde{W}^k$ in (5.38).
3. Update E^{k+1} by applying the soft-thresholding operator on $\widetilde{E}^k$ in (5.39).
4. Update $t_{k+1} = \frac{1}{2}(1 + \sqrt{1 + 4t_k^2})$.
5. $(\bar{W}^{k+1}, \bar{E}^{k+1}) = (W^{k+1}, E^{k+1}) + \frac{t_k - 1}{t_{k+1}}(W^{k+1} - W^k, E^{k+1} - E^k)$

until Convergence
Output: Accumulation points $\overline{W}$ and $\overline{E}$

As has already been proved in Beck and Teboulle [9], the APG algorithm, including the one above, has a very nice worst case iteration complexity result in that for any given $\epsilon > 0$, the APG algorithm needs at most $O(1/\sqrt{\epsilon})$ iterations to compute an ϵ-optimal (in terms of function value) solution.

The tuning of the regularization parameters λ and γ in (5.37) is fairly straightforward. For λ, we use the singular values of the converged $\overline{W}$ as a reference. Starting from a relatively large value of λ, we reduce it by a constant factor in each pass to obtain a $\overline{W}$ such that its singular values beyond the rth are much smaller than the first r singular values. For γ, we use the suggested value of $1/\sqrt{\max(m, n)}$ from RPCA [27]. In our experiments, we find that we only need a ballpark figure, without having to do a lot of tuning. Taking $\lambda = 0.1$ and $\gamma = 1/\sqrt{\max(m, n)}$ serve the purpose well.

5.4.6 Other heuristics

In practice, we design two heuristics to further boost the quality of the convex initialization. These are tricks that allow PARSuMi to detect corrupted entries better and are

always recommended.

We refer to the first heuristic as "Huber Regression". The idea is that the quadratic loss term in our matrix completion step (5.14) is likely to result in a dense spread of estimation error across all measurements. There is no guarantee that those true corrupted measurements will hold larger errors comparing to the uncorrupted measurements. On the other hand, we note that the quality of the subspace N^k obtained from LM_GN is usually good despite noisy/corrupted measurements. This is especially true when the first LM_GN step is initialized with Algorithm 5. Intuitively, we should be better off with an intermediate step, using N^{k+1} to detect the errors instead of W^{k+1}, that is, keeping N^{k+1} as a fixed input and finding coefficient C and E simultaneously with

$$\begin{aligned} &\underset{E,C}{\text{minimize}} \ \frac{1}{2}\|H \circ (N^{k+1}C - \widehat{W} + E)\|^2 \\ &\text{subject to} \quad \|E\|_0 \leq N_0. \end{aligned} \tag{5.40}$$

To make it computationally tractable, we relax (5.40) to

$$\underset{E,C}{\text{minimize}} \ \frac{1}{2}\|H \circ (N^{k+1}C - \widehat{W} + E)\|^2 + \eta_0\|E\|_1 \tag{5.41}$$

where $\eta_0 > 0$ is a penalty parameter. Note that each column of the above problem can be decomposed into the following Huber loss regression problem (E is absorbed into the Huber penalty)

$$\underset{C_j}{\text{minimize}} \sum_{i=1}^{m} \text{Huber}_{\eta_0/H_{ij}}(H_{ij}((N^{k+1}C_j)_i - \widehat{W}_{ij})). \tag{5.42}$$

Since N^{k+1} is known, (5.41) can be solved very efficiently using the APG algorithm, whose derivation is similar to that of Algorithm 5, with soft-thresholding operations on C and E. To further reduce the Robin Hood effect (that haunts all ℓ_1-like penalties) and enhance sparsity, we may optionally apply the iterative re-weighted Huber minimization (a slight variation of the method in Candes et al. [31]), that is, solving (5.42) for l_{max} iterations using an entrywise weighting factor inversely proportional to the previous

iteration's fitting residual. In the end, the optimal columns C_j's are concatenated into the optimal solution matrix C^* of (5.41), and we set

$$W^{k+1} = N^{k+1}C^*.$$

With this intermediate step between the W step and the E step, it is much easier for the E step to detect the support of the actual corrupted entries.

The above procedure can be used in conjunction with another heuristic that avoids adding false positives into the corruption set in the E step when the subspace N has not yet been accurately recovered. This is achieved by imposing a threshold η on the minimum absolute value of E^k's non-zero entries, and shrink this threshold by a factor (say 0.8) in each iteration. The "Huber regression" heuristic is used only when $\eta > \eta_0$, and hence only in a very small number of iteration before the support of E has been reliably recovered. Afterwards the pure PARSuMi iterations (without the Huber step) will take over, correct the Robin Hood effect of Huber loss and then converge to a high quality solution.

Note that our critical point convergence guarantee in Section 5.4.4 is not hampered at all by the two heuristics, since after a small number of iterations, $\eta \leq \eta_0$ and we come back to the pure PARSuMi.

5.5 Experiments and discussions

In this section, we present the methodology and results of various experiments designed to evaluate the effectiveness of our proposed method. The experiments revolve around synthetic data and two real-life datasets: the Oxford Dinosaur sequence, which is representative of data matrices in SfM works, and the Extended YaleB face dataset [91], which we use to demonstrate how PARSuMi works on photometric stereo problems.

In the synthetic data experiments, our method is compared with the state-of-the-art algorithms for the objective function in (5.10) namely Wiberg ℓ_1 [58] and GRASTA [71]. ALP and AQP [80] are left out since they are shown to be inferior to Wiberg ℓ_1

in Eriksson and Van Den Hengel [58]. For the sake of comparison, we perform the experiment on recovery effectiveness using the same small matrices as in Section 5.1 of Eriksson and Van Den Hengel [58]. Other synthetic experiments on Gaussian noise and phase diagram are conducted with more reasonably-sized matrices.

For the Dinosaur sequence, we investigate the quantitative effectiveness by adding *realistic large* errors to random locations of the data and checking against the known ground truth for E, and the qualitative effectiveness by looking at the trajectory plot which is revealing. We have normalized image pixel dimensions (width and height) to be in the range [0,1]; all plots, unless otherwise noted, are shown in the normalized coordinates. For the Extended YaleB, we reconstruct the full scale 3D face shape of all 38 subjects. Since there are no known locations for the corruption, we will carry out a qualitative comparison with the results of the nuclear norm minimization approach (first proposed in Wu et al. [149] to solve photometric stereo) and to BALM [46] which is a factorization method with specific manifold constraints for this problem.

Given the prevalence of convex relaxation of difficult problems in optimization, we also investigate the impact of convex relaxation as an initialization step. The fact that the initialization result is much less than desired also serves to vindicate our earlier statement about the relative merits of the nuclear norm minimization and the factorization approach.

In all our experiments, r is assumed to be known and N_0 is set to 1.2 times the true number of corruptions. In all synthetic data experiments, γ is fixed as $1/\sqrt{mn}$ for the initialization (5) and λ is automatically tuned using a binary search like algorithm to find a good point where the $(r+1)^{th}$ singular value of W is smaller than a threshold. In all real experiments, λ is set as 0.2. Our Matlab implementation is run on a 64-bit Windows machine with a 1.6 GHz Core i7 processor and 4 GB of memory.

5.5.1 Convex Relaxation as an Initialization Scheme

We first investigate the results of our convex initialization scheme by testing on a randomly generated 100×100 rank-4 matrix. A random selection of 70% and 10% of the

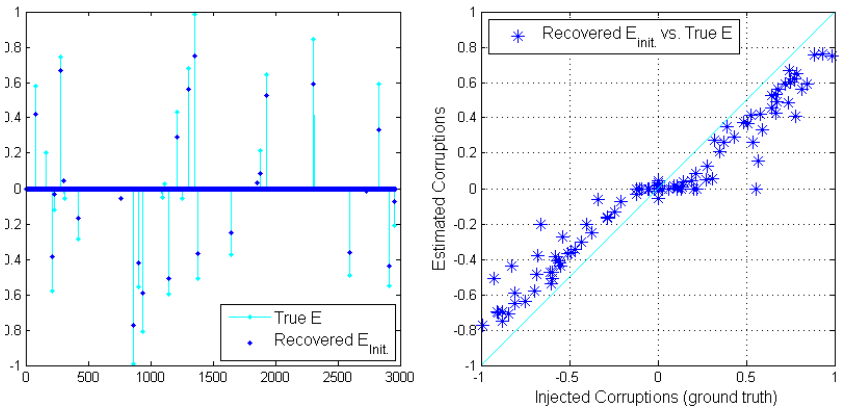


Figure 5.7: The Robin Hood effect of Algorithm 5 on detected sparse corruptions E_{Init}. **Left**: illustration of a random selection of detected E vs. true E. Note that the support is mostly detected, but the magnitude falls short. **Right**: scatter plot of the detected E against true E (perfect recovery falls on the $y = x$ line, false positives on the y-axis and false negatives on the x-axis).

entries are considered missing and corrupted respectively. Corruptions are generated by adding large uniform noise between $[-1, 1]$. In addition, Gaussian noise $N(0, \sigma)$ for $\sigma = 0.01$ is added to all observed entries. From Figure 5.7, we see that the convex relaxation outlined in Section 5.4.5 was able to recover the error support, but there is considerable difference in magnitude between the recovered error and the ground truth, owing to the "Robin Hood" attribute of ℓ_1-norm as a convex proxy of ℓ_0. Nuclear norm as a proxy of rank also suffers from the same woe, because nuclear norm and rank are essentially the ℓ_1 and ℓ_0 norm of the vector of singular values respectively. As clearly illustrated in Figure 5.8, the recovered matrix from Algorithm 5 has smaller first four singular values and non-zero singular values beyond the fourth. Similar observations can be made on the results of the Dinosaur experiments, which we will show later.

Despite the problems with the solution of the convex initialization, we find that it is a crucial step for PARSuMi to work well in practice. As we have seen from Figure 5.7, the detected error support can be quite accurate. This makes the E-step of PARSuMi more likely to identify the true locations of corrupted entries.

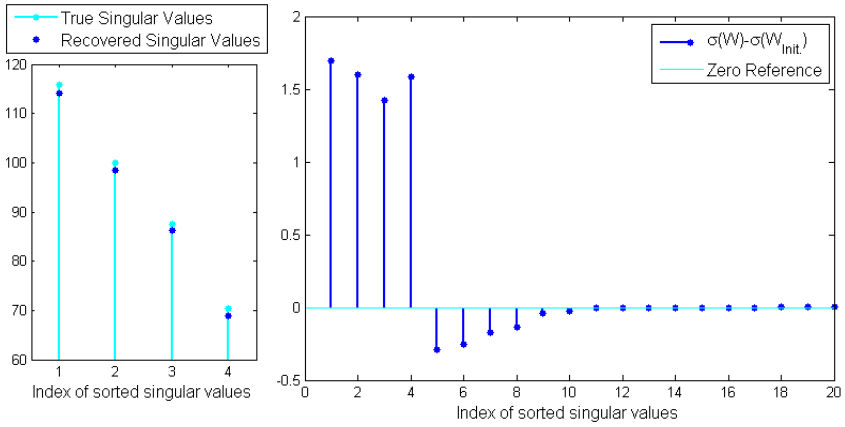
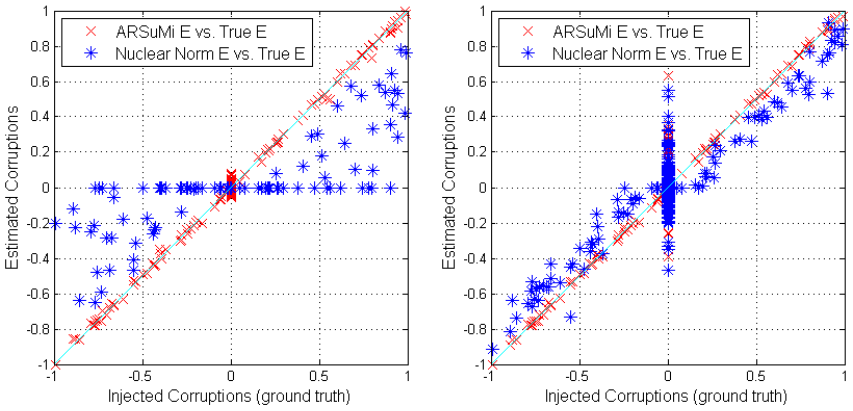


Figure 5.8: The Robin Hood effect of Algorithm 5 on singular values of the recovered $W_{\text{Init.}}$. **Left**: illustration of the first 4 singular values. Note that the magnitude is smaller than that of the ground truth. **Right**:The difference of the true and recovered singular values (first 20). Note that the first 4 are positive and the rest are negative.

5.5.2 Impacts of poor initialization

When the convex initialization scheme fails to obtain the correct support of the error, the "Huber Regression" heuristic may help PARSuMi to identify the support of the corrupted entries. We illustrate the impact by intentionally mis-tuning the parameters of Algorithm 5 such that the initial E bears little resemblance to the true injected corruptions. Specifically, we test the cases when the initialization fails to detect many of the corrupted entries (false negatives) and when many entries are wrongly detected as corruptions (false positives). From Figure 5.9, we see that PARSuMi is able to recover the corrupted entries to a level comparable to the magnitude of the injected Gaussian noise in both experiments. Note that a number of false positives persist in the second experiment. This is understandable because false positives often contaminate an entire column or row, making it impossible to recover that column/row in later iterations even if the subspace is correctly detected[1]. To avoid such an undesirable situation, we prefer "false negatives" over "false positives" when tuning Algorithm 5. In practice, it suffices to keep the initial E relatively sparse.

[1]We may add arbitrary error vector in the span of the subspace. In the extreme case, all observed entries in a column can be set to zero.



(a) False negatives (b) False positives

Figure 5.9: Recovery of corruptions from poor initialization.

In most of our experiments, we find that PARSuMi is often able to detect the corruptions perfectly from a simple initializations with all zeros, even without the "Huber Regression" heuristic. This is especially true when the data are randomly generated with benign sampling pattern and well-conditioned singular values. However, in challenging applications such as SfM, a good convex initialization and the "Huber Regression" heuristic are always recommended.

5.5.3 Recovery effectiveness from sparse corruptions

For easy benchmarking, we use the same synthetic data in Section 5.1 of Eriksson and Van Den Hengel [58] to investigate the quantitative effectiveness of our proposed method. A total of 100 random low-rank matrices with missing data and corruptions are generated and tested using PARSuMi, Wiberg ℓ_1 and GRASTA.

In accordance with Eriksson and Van Den Hengel [58], the ground truth low rank matrix $W_{\text{groundtruth}} \in \mathbb{R}^{m\times n}, m = 7, n = 12, r = 3$, is generated as $W_{\text{groundtruth}} = UV^T$, where $U \in \mathbb{R}^{m\times r}, V \in \mathbb{R}^{n\times r}$ are generated using uniform distribution, in the range [-1,1]. 20% of the data are designated as missing, and 10% are added with corruptions, both at random locations. The magnitude of the corruptions follows a uniform distribution $[-5, 5]$. Root mean square error(RMSE)is used to evaluate the recovery

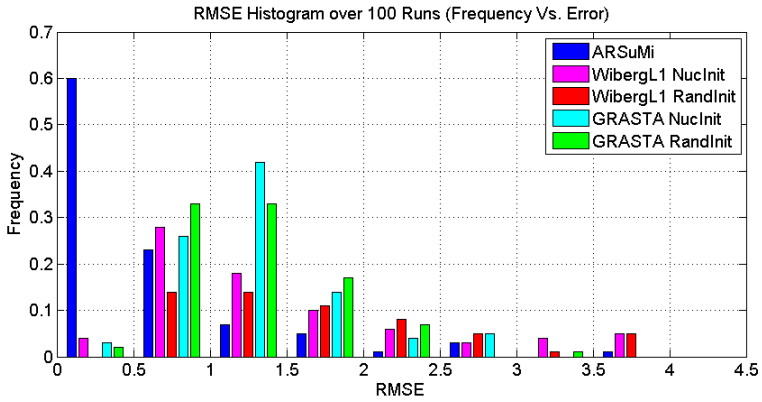


Figure 5.10: A histogram representing the frequency of different magnitudes of RMSE in the estimates generated by each method.

precision:

$$\text{RMSE} := \frac{\|W_{\text{recovered}} - W_{\text{groundtruth}}\|_F}{\sqrt{mn}}. \tag{5.43}$$

Out of the 100 independent experiments, the number of runs that returned RMSE values of less than 5 are 100 for PARSuMi, 78 and 58 for Wiberg ℓ_1 (with two different initializations) and similarly 94 and 93 for GRASTA. These are summarized in Figure 5.10. We see that our method has the best performance. Wiberg ℓ_1 and GRASTA performed similarly, though GRASTA converged to a reasonable solution more often. In addition, our convex initialization improves the results of Wiberg ℓ_1 and GRASTA, though not significantly.

5.5.4 Denoising effectiveness

An important difference between our method and the algorithms that solve (5.10) (e.g., Wiberg ℓ_1) is the explicit modelling of Gaussian noise. We set up a synthetic rank-4 data matrix of size 40×60, with 50% missing entries, 10% sparse corruptions in the range $[-5, 5]$, and Gaussian noise $\mathcal{N}(0, \sigma)$ with standard deviation σ in the range [0,0.2]. The amount of missing data and corruptions are selected such that both Wiberg ℓ_1 and PARSuMi can confidently achieve exact recovery in noise-free scenario. We also adapt the oracle lower bound from Candes and Plan [21] to represent the theoretical limit of re-

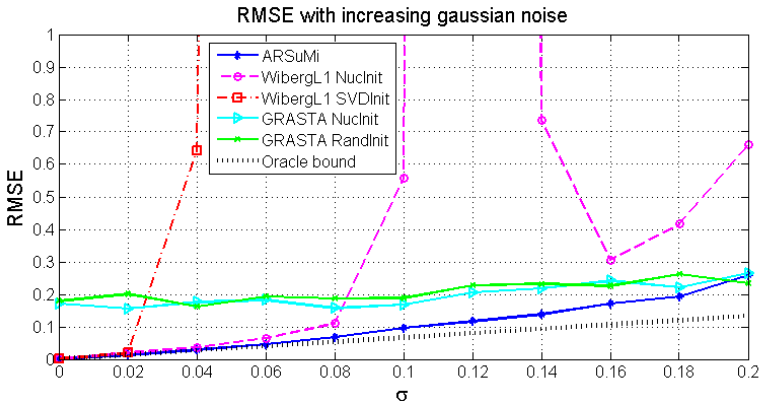


Figure 5.11: Effect of increasing Gaussian noise: PARSuMi is very resilient while Wiberg ℓ_1 becomes unstable when noise level gets large. GRASTA is good when noise level is high, but not able to converge to a good solution for small σ even if we initialize it with Algorithm 5.

covery accuracy under noise. Our extended oracle bound under both sparse corruptions and Gaussian noise is:

$$\text{RMSE}_{oracle} = \sigma\sqrt{\frac{(m+n-r)r}{p-e}}, \tag{5.44}$$

where p is the number of observed entries and e is the number of corruptions in the observations.

We see from Figure 5.11 that under such conditions, Wiberg ℓ_1 is able to tolerate small Gaussian noise, but becomes unstable when the noise level gets higher. In contrast, since our method models Gaussian noise explicitly, the increasing noise level has little impact. In particular, our performance is close to the oracle bound. Moreover, we observe that GRASTA is not able to achieve a high quality recovery when the noise level is low, but becomes near optimal when σ gets large.

Another interesting observation is that Wiberg ℓ_1 with convex relaxation as initialization is more tolerant to the increasing Gaussian noise. This could be due to the better initialization, since the convex relaxation formulation also models Gaussian noise.

5.5.5 Recovery under varying level of corruptions, missing data and noise

The experiments conducted so far investigate only specific properties. To gain a holistic understanding of our proposed method, we perform a series of systematically parameterized experiments on 40×60 rank-4 matrices (with the elements of the factors U, V drawn independently from the uniform distribution on $[-1, 1]$), with conditions ranging from 0-80% missing data, 0-20% corruptions of range [-2,2], and Gaussian noise with σ in the range [0,0.1]. By fixing the Gaussian noise at a specific level, the results are rendered in terms of phase diagrams showing the recovery precision as a function of the missing data and outliers. The precision is quantified as the difference between the recovered RMSE and the oracle bound RMSE. As can be seen from Figure 5.12(a), our algorithm obtains near optimal performance at an impressively large range of missing data and outlier at $\sigma = 0.01$[1].

For comparison, we also displayed the phase diagram of our convex initialization in Figure 5.12(b) and that for GRASTA from two different initialization schemes in Figure 5.12(c) and 5.12(d), Wiberg ℓ_1 is omitted because it is too slow. Without the full non-convex machinery, the relaxed version is not able to reconstruct the exact matrix. Its RMSE value grows substantially with the increase of missing data and outliers. GRASTA is also incapable of denoising and cannot achieve a high-precision recovery result even when there is neither missing nor corrupted data (at the top left corner).

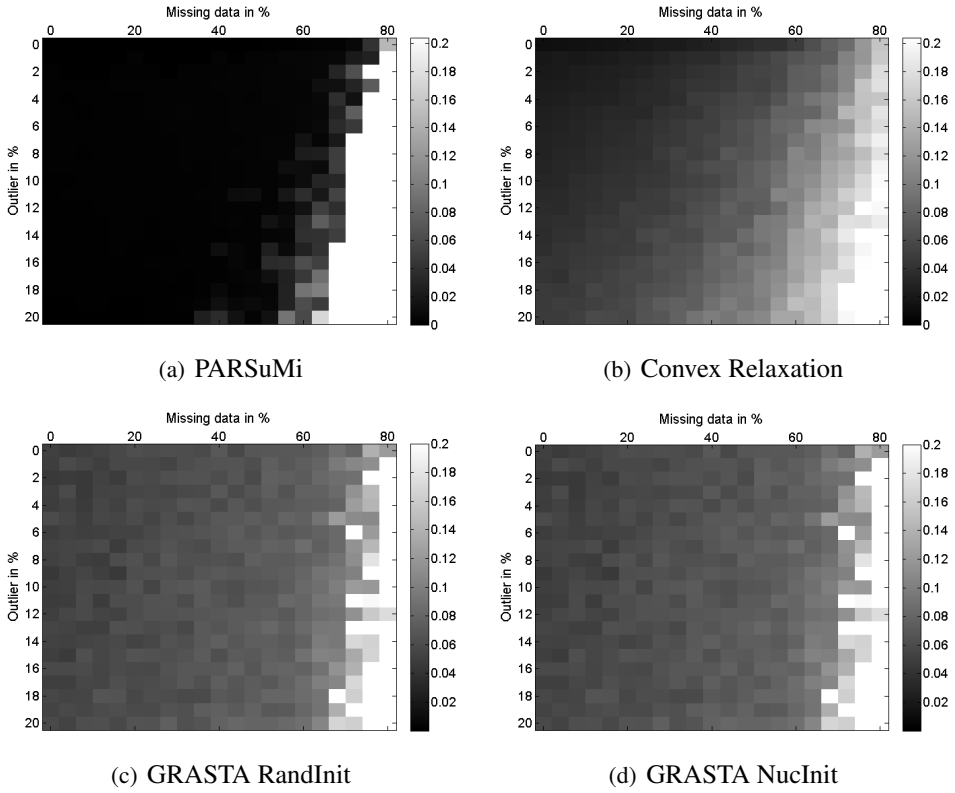
5.5.6 SfM with missing and corrupted data on Dinosaur

In this section, we apply PARSuMi to the problem of SfM using the Dinosaur sequence and investigate how well the corrupted entries can be detected and recovered in real data.

To simulate data corruptions arising from wrong feature matches, we randomly add sparse error of the range [-2,2][2] to 1% of the sampled entries. This is a more realistic

[1]The phase diagrams for other levels of noise look very much like Figure 5.12; we therefore did not include them.

[2]In SfM data corruptions are typically matching failures. Depending on where true matches are, error induced by a matching failure can be arbitrarily large. If we constrain true match to be inside image frame



(a) PARSuMi

(b) Convex Relaxation

(c) GRASTA RandInit

(d) GRASTA NucInit

Figure 5.12: Phase diagrams (darker is better) of RMSE with varying proportion of missing data and corruptions with Gaussian noise $\sigma = 0.01$ (roughly 10 pixels in a 1000×1000 image).

(and much larger[1]) definition of outliers for SfM compared to the [-50,50] pixel range used to evaluate Wiberg ℓ_1 in Eriksson and Van Den Hengel [58]. In fact, both algorithms work almost perfectly under the condition given in Section 5.2 of Eriksson and Van Den Hengel [58]. An evaluation on larger corruptions helps to show the differing performance under harsher condition.

We conducted the experiment 10 times each for PARSuMi, Wiberg ℓ_1 (with SVD initialization) and GRASTA (random initialization as recommended in the original paper) and count the number of times they succeed. As there are no ground truth to com-

$[0, 1]$(which is often not the case), then the maximum error magnitude is 1. We found it appropriate to at least double the size to account for general matching failures in SfM, hence $[-2, 2]$.

[1][-50,50] in pixel is only about [-0.1,0.1] in our normalized data, which could hardly be regarded as "gross" corruptions.

	PARSuMi	Wiberg ℓ_1	GRASTA
No. of success	9/10	0/10	0/10
Run time (mins): min/avg/max	2.2/2.9/5.2	76/105/143	0.2/0.5/0.6
Min RMSE (original pixel unit)	1.454	2.715	22.9
Min RMSE excluding corrupted entries	0.3694	1.6347	21.73

Table 5.3: Summary of the Dinosaur experiments. Note that because there is no ground truth for the missing data, the RMSE is computed only for those observed entries as in Buchanan and Fitzgibbon [19].

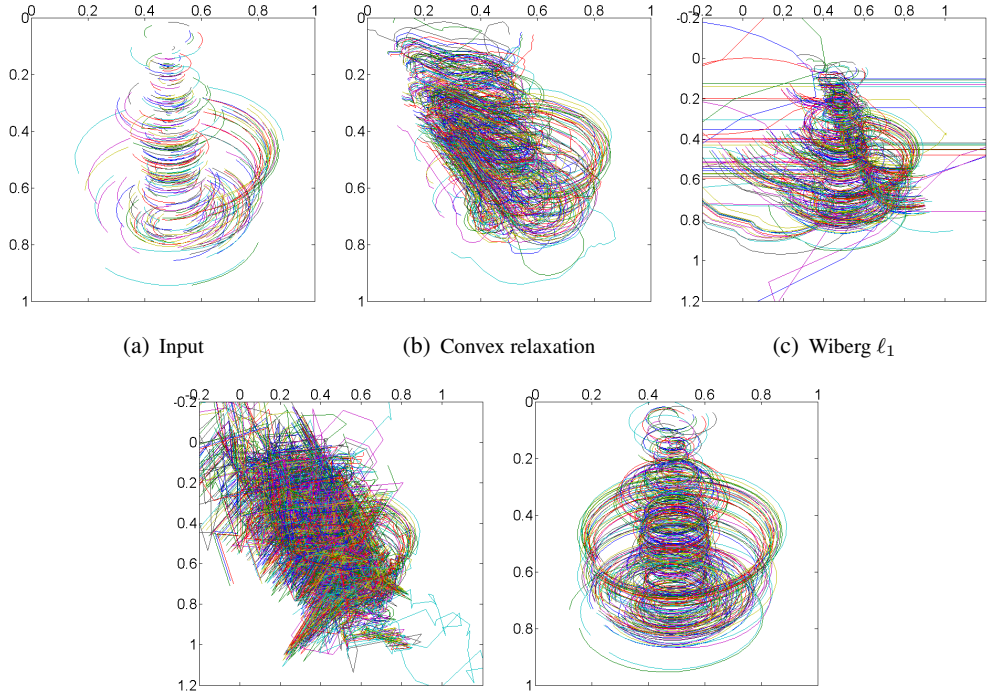
pare against, we cannot use the RMSE to evaluate the quality of the filled-in entries. Instead, we plot the feature trajectory of the recovered data matrix for a qualitative judgement. As is noted in Buchanan and Fitzgibbon [19], a correct recovery should consist of all elliptical trajectories. Therefore, if the recovered trajectories look like that in Figure 5.6(b), we count the recovery as a success.

The results are summarized in Table 5.3. Notably, PARSuMi managed to correctly detect the corrupted entries and fill in the missing data in 9 runs while Wiberg ℓ_1 and GRASTA failed on all 10 attempts. Typical feature trajectories recovered by each method are shown in Figure 5.13. Note that only PARSuMi is able to recover the elliptical trajectories satisfactorily.

For comparison, we also include the input (partially observed trajectories) and the results of our convex initialization in Figure 5.13(a) and 5.13(b) respectively. Due to the Robin Hood attribute of nuclear norm, the filled-in trajectories of the convex relaxation has a significant bias towards smaller values (note that the filled-in shape tilts towards the origin). This is because nuclear norm is not as magnitude insensitive as the rank function. Smaller filled-in data usually lead to a smaller nuclear norm.

Another interesting and somewhat surprising finding is that the result of PARSuMi is even better than the global optimal solution for data containing supposedly no corruptions (and thus can be obtained with ℓ_2 method) (see Figure 5.6(b), which is obtained under no corruptions in the observed data)! In particular, the trajectories are now closed.

The reason becomes clear when we look at Figure 5.14(b), which shows two large spikes in the vectorized difference between the artificially injected corruptions and the



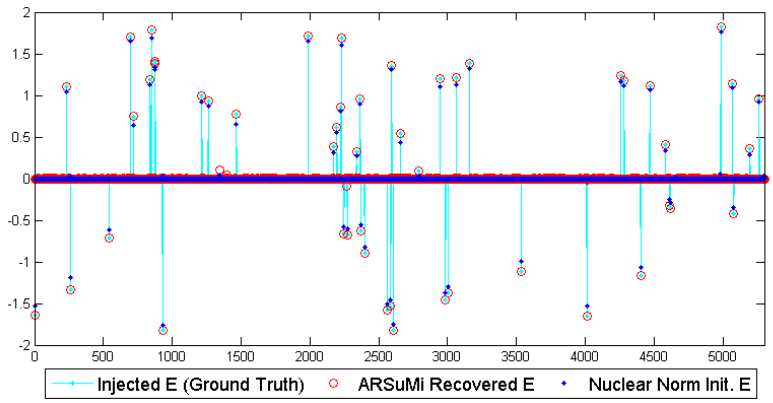
(a) Input (b) Convex relaxation (c) Wiberg ℓ_1

(d) GRASTA (e) PARSuMi

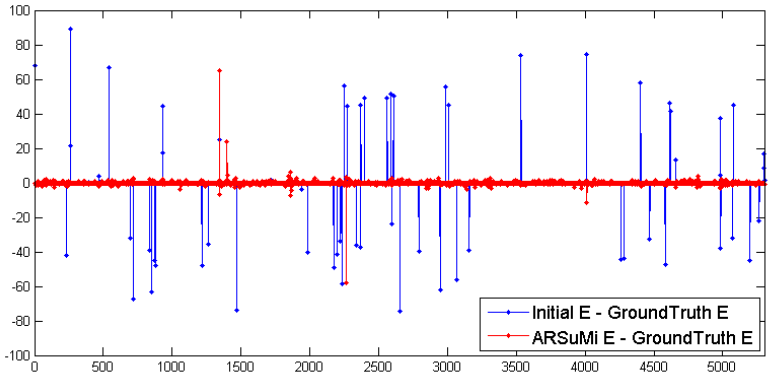
Figure 5.13: Comparison of recovered feature trajectories with different methods. It is clear that under dense noise and gross outliers, neither convex relaxation nor ℓ_1 error measure yields satisfactory results. Solving the original non-convex problem with (b) as an initialization produces a good solution.

recovered corruptions by PARSuMi. This suggests that there are hitherto unknown corruptions inherent in the Dinosaur data. We trace the two large ones into the raw images, and find that they are indeed data corruptions corresponding to mismatched feature points from the original dataset (see Figure 5.15); our method managed to recover the correct feature matches (left column of Figure 5.15).

The result shows that PARSuMi recovered not only the artificially added errors, but also the intrinsic errors in the data set. In Buchanan and Fitzgibbon [19], it was observed that there is a mysterious increase of the objective function value upon closing the trajectories by imposing orthogonality constraint on the factorized camera matrix. Our discovery of these intrinsic tracking errors explained this matter evidently. It is also the reason why the ℓ_2-based algorithms find a global minimum solution that is of poorer



(a) Initialization via Algorithm 5 and the final recovered errors by PARSuMi (Algorithm 3)



(b) Difference of the recovered and ground truth error (in original pixel unit)

Figure 5.14: Sparse corruption recovery in the Dinosaur experiments: The support of all injected outliers are detected by Algorithm 5 (see (a)), but the magnitudes fall short by roughly 20% (see (b)). Algorithm 3 is able to recover all injected sparse errors, together with the inherent tracking errors in the dataset (see the red spikes in (b)).

quality (trajectories fail to close loop).

To complete the story, we generated the 3D point cloud of Dinosaur with the completed data matrix. The results viewed from different directions are shown in Figure 5.16.

5.5.7 Photometric Stereo on Extended YaleB

Another intuitive application for PARSuMi is photometric stereo, a problem of reconstructing the 3D shape of an object from images taken under different lighting con-

(a) Zoom-in view: recovered matching error in frame 13

(b) Zoom-in view: recovered matching error in frame 15

Figure 5.15: Original tracking errors in the Dinosaur data identified (yellow box) and corrected by PARSuMi (green box with red star) in frame 13 feature 86 (a) and frame 15 feature 144 (b).

ditions. In the most ideal case of Lambertian surface model (diffused reflection), the intensity of each pixel is proportional to the inner product of the surface normal $\vec{n}$ associated with the pixel and the lighting direction $\vec{l}$ of the light source. This leads to the matrix factorization model

$$[I_1, ..., I_k] = \rho \begin{pmatrix} \alpha_1 \vec{n_1}^T \\ ... \\ \alpha_p \vec{n_p}^T \end{pmatrix} \begin{pmatrix} L_1 \vec{l_1} & ... & L_k \vec{l_k} \end{pmatrix} = \rho A^T B, \tag{5.45}$$

where I_j represents the vectorized greyscale image taken under lighting j, ρ is the Lambertian coefficient, α_i is the albedo of a pixel i, and L_j is the light intensity in image j. The consequence is that the data matrix obtained by concatenating vectorized images together is of rank 3.

Real surfaces are of course never truly Lambertian. There are usually some localized specular regions appearing as highlights in the image. Moreover, since there is no way to obtain a negative pixel value, all negative inner products will be observed as zero. This is the so-called attached shadow. Images of non-convex object often also contain cast shadow, due to the blocking of light path. If these issues are teased out, then the seemingly naive Lambertian model is able to approximate many surfaces very well.

Wu et al. [149] subscribed to this low-rank factorization model in (5.45) and proposed to model all dark regions as missing data, all highlights as sparse corruptions and

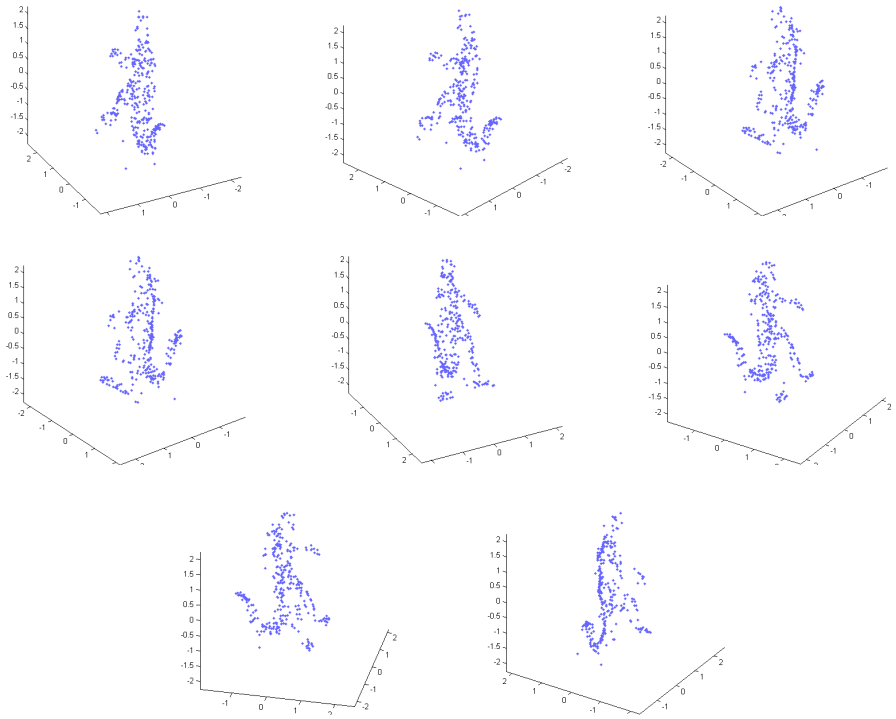


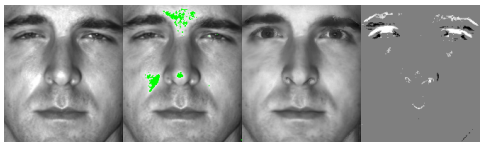
Figure 5.16: 3D point cloud of the reconstructed Dinosaur.

then use a variant of RPCA (identical to (5.6)) to recover the full low-rank matrix. The solution however is only tested on noise-free synthetic data and toy-scale real examples. Del Bue et al. [46] applied their BALM on photometric stereo too, attempting on both synthetic and real data. Their contribution is to impose the normal constraint of each normal vector during the optimization. Del Bue et. al. also propose using a sophisticated inpainting technique to initialize the missing entries in the image, which is likely to improve the chance of BALM converging to a good solution. Later we will provide a qualitative comparison of the results obtained by BALM, our convex initialization and PARSuMi. Note that the method in Wu et al. [149] is almost the same as our initialization, except that it does not explicitly handle noise. Since they have not released their source code, we will simply use Algorithm 5 to demonstrate the performance of this type of convex relaxation methods.

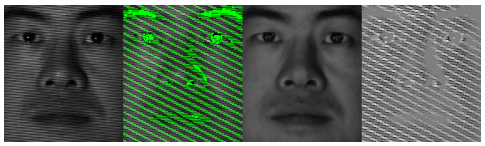
Methodology: To test the effectiveness of PARSuMi on full scale real data, we run

through all 38 subjects in the challenging Extended YaleB face database. The data matrix for each subject is a 32256×64 matrix where each column represents a vectorized $x \times y$ image and each row gives the intensities of a particular pixel across all 64 lighting conditions. After setting the shadow and highlight as missing data by thresholding[1], about 65% of the data are observed, with the sampling distribution being rather skewed (for some images, only 5-10% of the pixels are measured). In addition, subjects tend to change facial expressions in different images and there are some strange corruptions in the data, hence jeopardizing the rank-3 assumption. We model these unpredictable issues as sparse corruptions.

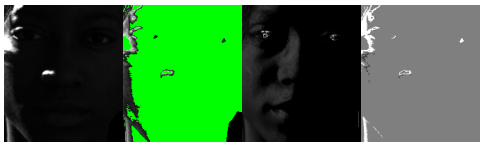
(a) Cast shadow and attached shadow are recovered. Region of cast shadow is now visible, and attached shadow is also filled with meaningful negative values.



(b) Facial expressions are set to normal.



(c) Rare corruptions in image acquisition are recovered.

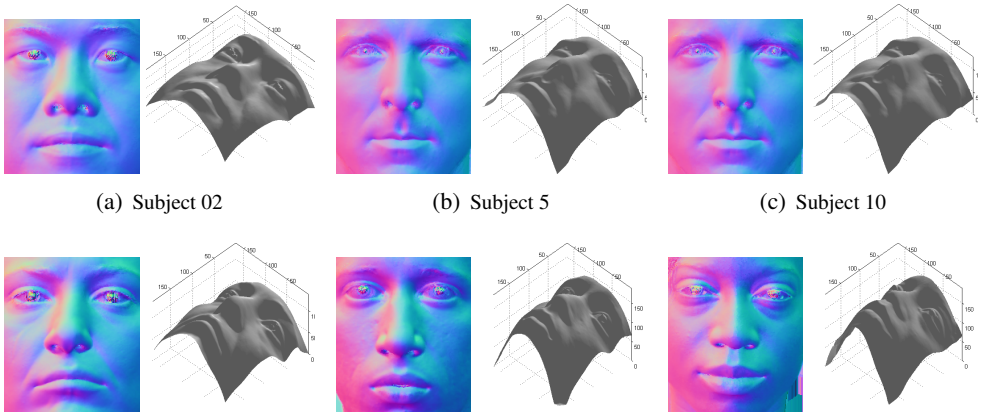


(d) Light comes from behind (negative 20 degrees to the horizontal axis and 65 degrees to the vertical axis).

Figure 5.17: Illustrations of how PARSuMi recovers missing data and corruptions. From left to right: original image, input image with missing data labeled in green, reconstructed image and detected sparse corruptions.

Results: PARSuMi is able to successfully reconstruct the 3D face of all 38 subjects with little artifacts. An illustration of the input data and how PARSuMi recovers the missing elements and corruptions are shown in Figure 5.17, and the reconstruction of selected faces across genders and ethnic groups are shown in Figure 5.18. We remark that the results are of high precision and even the tiny wrinkles and moles on the faces can be observed. Furthermore, we attach the results of all 64 images of Subject 10 in the Appendix (Figure D.1) for further scrutiny by interested readers.

[1] In our experiment, all pixels with values smaller than 20 or greater than 240 are set as missing data.



(a) Subject 02 (b) Subject 5 (c) Subject 10

(d) Subject 15 (e) Subject 12 (f) Subject 22

Figure 5.18: The reconstructed surface normal and 3D shapes for Asian (first row), Caucasian (second row) and African (third row), male (first column) and female (second column), in Extended YaleB face database.(Zoom-in to look at details)

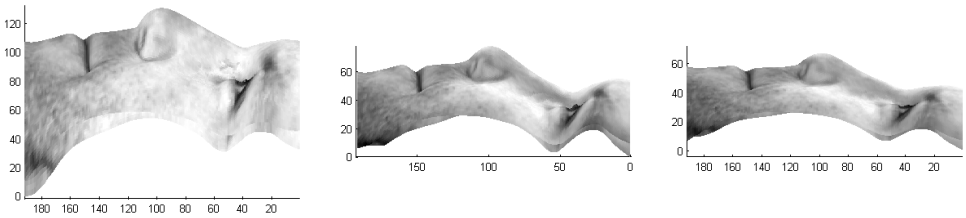
We compare PARSuMi, BALM and our convex initialization using Subject 3 in the YaleB dataset since it was initially used to evaluate BALM in Del Bue et al. [46][1]. The results are qualitatively compared in Figure 5.19. As we can see, both BALM and Algorithm 5 returned obvious artifact in the recovered face image, while PARSuMi's results looked significantly better. The difference manifests itself further when we take the negative of the recovered images by the three algorithms (see Figure 5.19(c)). From (5.45), it is clear that taking negative is equivalent to inverting the direction of lighting. The original lighting is -20° from the left posterior and 40° from the top, so the inverted light should illuminate the image from the right and from below. The results in Figure 5.19(c) clearly show that neither BALM nor Algorithm 5 is able to recover the missing data as well as PARSuMi. In addition, we reconstruct the 3D depth map with the classic method by Horn [73] and show the side face in Figure 5.19(d). The shape from PARSuMi reveals much richer depth information than those from the other two algorithms, whose reconstructions appear flattened.

[1]The authors claimed that it is Subject 10 [46, Figure 9], but careful examination of all faces shows that it is in fact Subject 3.

(a) Comparison of the recovered image

(b) Comparison of the recovered image (details)

(c) Taking the negative of (a) to see the filled-in negative values.



(d) Comparison of reconstructed 3D surface (albedo rendered)

Figure 5.19: Qualitative comparison of algorithms on Subject 3. From left to right, the results are respectively for PARSuMi, BALM and our convex initialization. In (a) and (c), they are preceded by the original image and the image depicting the missing data in green.

5.5.8 Speed

The computational complexity of PARSuMi is cheap for some problems but not for others. Since PARSuMi uses LM_GN for its matrix completion step, the numerical cost is dominated by either solving the linear system $(J^TJ + \lambda I)\delta = J\mathbf{r}$ which requires the Cholesky factorization of a potentially dense $mr \times mr$ matrix, or the computation of J which requires solving a small linear system of normal equation involving the $m \times r$ matrix N for n times. As the overall complexity of $O(\max(m^3r^3, mnr^2))$ scales merely linearly with number of columns n but cubic with m and r, PARSuMi

is computationally attractive when solving problems with small m and r, and large n, e.g., photometric stereo and SfM (since the number of images is usually much smaller than the number of pixels and feature points). However, for a typical million by million data matrix as in social networks and collaborative filtering, PARSuMi will take an unrealistic amount of time to run.

Experimentally, we compare the runtime between our algorithm and Wiberg ℓ_1 method in our Dinosaur experiment in Section 5.5.6. We see from Table 5.3 that there is a big gap between the speed performance. The near 2-hour runtime for Wiberg ℓ_1 is discouragingly slow, whereas ours is vastly more efficient. On the other hand, as an online algorithm, GRASTA is inherently fast. Examples in He et al. [71] show that it works in real time for live video surveillance. However, our experiment suggests that it is probably not appropriate for applications such as SfM, which requires a higher numerical accuracy.

We note that PARSuMi is currently not optimized for computation. Speeding up the algorithm for application on large scale dataset would require further effort (such as parallelization) and could be a new topic of research. For instance, the computation of Jacobians J_i and evaluating objective function can be easily done in parallel and the Gauss-Newton update (a positive definite linear system of equations) can be solved using the conjugate gradient method; hence, we do not even need to store the matrix in memory. Furthermore, since PARSuMi seeks to find the best subspace, perhaps using only a small portion of the data columns is sufficient. If the subspace is correct, the rest of the columns can be recovered in linear time with our iterative reweighted Huber regression technique (see Section 5.4.6). A good direction for future research is perhaps on how to choose the best subset of data to feed into PARSuMi.

5.6 Chapter Summary

In this chapter, we have presented a practical algorithm (PARSuMi) for low-rank matrix completion in the presence of dense noise and sparse corruptions. Despite the

non-convex and non-smooth optimization formulation, we are able to derive a set of update rules under the proximal alternating scheme such that the convergence to a critical point can be guaranteed. The method was tested on both synthetic and real life data with challenging sampling and corruption patterns. The various experiments we have conducted show that our method is able to detect and remove gross corruptions, suppress noise and hence provide a faithful reconstruction of the missing entries. By virtue of the explicit constraints on both the matrix rank and cardinality, and the novel reformulation, design and implementation of appropriate algorithms for the non-convex and non-smooth model, our method works significantly better than the state-of-the-art algorithms in nuclear norm minimization, ℓ_2 matrix factorization and ℓ_1 robust matrix factorization in real life problems such as SfM and photometric stereo.

Moreover, we have provided a comprehensive review of the existing results pertaining to the "practical matrix completion" problem that we considered in this chapter. The review covered the theory of matrix completion and corruption recovery, and the theory and algorithms for matrix factorization. In particular, we conducted extensive numerical experiments which reveals (a) the advantages of matrix factorization over nuclear norm minimization when the underlying rank is known, and (b) the two key factors that affect the chance of ℓ_2-based factorization methods reaching global optimal solutions, namely "subspace parameterization" and "Gauss-Newton" update. These findings provided critical insights into this difficult problem, upon the basis which we developed PARSuMi as well as its convex initialization.

The strong empirical performance of our algorithm calls for further analysis. For instance, obtaining the theoretical conditions for the convex initialization to yield good support of the corruptions should be plausible (following the line of research discussed in Section 5.2.1), and this in turn guarantees a good starting point for the algorithm proper. Characterizing how well the following non-convex algorithm works given such initialization and how many samples are required to guarantee high-confidence recovery of the matrix remain open questions for future study.

Other interesting topics include finding a cheaper but equally effective alternative

to the LM_GN solver for solving (5.20), parallel/distributed computation, incorporating additional structural constraints, selecting optimal subset of data for subspace learning and so on. Step by step, we hope this will eventually lead to a practically working robust matrix completion algorithm that can be confidently embedded in real-life applications.

Chapter 6

Conclusion and Future Work

This thesis investigates the problem of robust learning with two prevalent low-dimensional structures: low-rank subspace model and the union-of-subspace model. The results are encouraging in both theoretical and algorithmic fronts. With the well-justified robustness guarantee, the techniques developed in this thesis can often be directly applied to real problems, even under considerable noise and model inaccuracy. In this chapter, we briefly summarize the contribution of the thesis and then list the open questions for future research.

6.1 Summary of Contributions

In Chapter 2 and 3, we considered two empirically working yet theoretically unsupported methods, matrix factorization and the noisy variant of SSC. By rigorous analysis of each method with techniques in compressive sensing, convex optimization, and statistical learning theory, we are able to understand their behaviors under noise/perturbations hence justify their good performance on real data. Furthermore, the results clearly identifies the key features of the problems that can be robustly solved and those that are more sensitive to noise thereby providing guidelines to practitioners, in particular, in designing collaborative filtering systems or doing clustering analysis of high dimensional data. In the context of machine learning, the main result in Chapter 2 can be

considered a generalization bound with natural implication on sample complexity (how many iid observations are needed).

In Chapter 4, we proposed a method that build upon the two arguably most successful subspace clustering methods (LRR and SSC). We demonstrated that their advantages can be combined but not without some tradeoff. The ℓ_1 penalty induces sparsity not only between classes but also within each class, while the nuclear norm penalty in general promotes a dense connectivity in both instances too. Interestingly, the analysis suggests that perfect separation can be achieved whenever the weight on ℓ_1 norm is greater than the threshold, thus showing that the best combination in practice is perhaps not the pure SSC or LRR but is perhaps a linear combination of them, i.e., LRSSC.

In Chapter 5, we focused on modelling and corresponding non-convex optimization for the so-called "Practical Matrix Completion" problem. It is related to Chapter 2 in that it seeks to solve a fixed rank problem. The problem is however much harder due to the possible gross corruptions in data. Our results suggest that the explicit modelling of PARSuMi provides substantial advantages in denoising, corruption recovery and in learning the underlying low-rank subspace over convex relaxation methods. At a point where the nuclear norm and ℓ_1-norm approaches are exhausting their theoretical challenges and reaching a bottleneck in practical applications, it may be worthwhile for the field to consider an alternate path.

6.2 Open Problems and Future Work

The works in this thesis also point to a couple of interesting open problems. These could be the future directions of research.

Theoretical foundation for matrix factorization

We studied the robustness of matrix factorization in Chapter 2 and showed that its global optimal solution has certain desirable properties, the most daunting problem: under what conditions the global optimal solution can be obtained and how to obtain it is still

an open question.

As a start, Jain et al. [76] analyzed the performance of ALS for the noiseless case, but there is an apparent gap between their assumptions and what empirical experiments showed. In particular, our evaluation in Section 5.3 suggests that ALS may not be the best approach for solving MF. Further improvement on the conditions in [76] and to allow for noise are clearly possible and should reveal good tricks to improve the performance of MF in practical problems.

Graph connectivity and missing data in subspace clustering

For the problem of subspace clustering, our results for LRSSC guarantee self-expressiveness at points when the solution is intuitively and empirically denser than SSC, yet there is still a gap in quantifying the level of connection density and in showing how dense a connectivity would guarantee that each block is a connected-body.

Missing data is another problem for subspace clustering techniques that exploit the intuition of self-expressiveness (SSC and LRR). A sampling mask like matrix completion in the constraint essentially makes the problem non-convex. Eriksson et al. [59] proposed the first provable algorithm for the missing data problem in subspace clustering using a bottom-up nearest neighbor-based approach, however require an unrealistic number of samples for each subspace, which could hardly be met in practice due to time and budget constraints. Advances on this missing data problem could potentially lead to immediate applications in the community clustering of social networks and motion segmentation in computer vision.

What we find interesting in this thesis is that we can use the same techniques (with minor adaptations) for simple structures like low-rank and sparsity to devise solutions for more sophisticated structures such as union-of-subspace model. Therefore, the key elements for solving the connectivity problem and missing data problem are probably already out there in the literature awaiting discovery.

General manifold clustering problem

From a more general point of view, the subspace clustering problem can be considered a special case of the manifold clustering problem. Is is possible to provably cluster data on general manifolds using the same intuition of "self-expressiveness" and with convex optimization[1]? On the other hand, could the rich topological structures of some manifolds (see [5]) be exploited in the problem of clustering?

This direction may potentially result in a uniform theory of clustering and unsupervised learning and go well beyond the current solutions such as k-means and spectral clustering [5].

Scalability for the big data: algorithmic challenges

As we have shown in this thesis, exploiting the low-dimensional structures is the key to gain statistical tractability for big and high dimensional data. It remains a computational challenge to actually solve these structure learning problems for internet-scale data in a reasonable amount of time.

Proposals such as matrix completion/RPCA as well as Lasso-SSC and LRSSC introduced in this thesis are typically just a convex optimization formulation. While one can be solved them in polynomial time with off-the-shelf SDP solvers, large-scale applications which often requires linear or even sub-linear runtime. Our proposed numerical solvers for our methods (ADMM algorithms for Matrix-Lasso-SSC in Chapter 3 and LRSSC in Chapter 4) could scale up for data matrices in the scale of tens of thousands, but is still considered impropriate for problems in the scale of millions and billions as described in the very beginning of this thesis.

It is therefore essential to adopt techniques such as divide-and-conquer for batch processing and incremental updates that minimizes memory cost. The algorithmic and theoretical challenge is to design large-scale extensions that can preserve the robustness and other good properties of the original methods. Results in this front will naturally attract avid attention in the emerging data industry.

[1] Elhamifar and Vidal [55] explored this possibility with some empirical results.

References

[1] M. Aharon, M. Elad, and A. Bruckstein. K-svd: Design of dictionaries for sparse representation. *Proceedings of SPARS*, 5:9–12, 2005. 4

[2] D. Alonso-Gutiérrez. On the isotropy constant of random convex sets. *Proceedings of the American Mathematical Society*, 136(9):3293–3300, 2008. 44, 55, 206

[3] H. Attouch, J. Bolte, P. Redont, and A. Soubeyran. Proximal alternating minimization and projection methods for nonconvex problems: An approach based on the kurdyka-łojasiewicz inequality. *Mathematics of Operations Research*, 35(2):438–457, 2010. 77, 92, 100, 102, 103

[4] Y. Azar, A. Fiat, A. Karlin, F. McSherry, and J. Saia. Spectral analysis of data. In *STOC*, pages 619–626, 2001. 10

[5] S. Balakrishnan, A. Rinaldo, D. Sheehy, A. Singh, and L. Wasserman, editors. *The NIPS 2012 Workshop on Algebraic Topology and Machine Learning.*, Lake Tahoe, Nevada, Dec. 2012. 132

[6] K. Ball. An elementary introduction to modern convex geometry. *Flavors of geometry*, 31:1–58, 1997. 175, 207

[7] L. Balzano, R. Nowak, and B. Recht. Online identification and tracking of subspaces from highly incomplete information. In *Communication, Control, and Computing (Allerton), 2010 48th Annual Allerton Conference on*, pages 704–711. IEEE, 2010. 83, 217

[8] R. Basri and D. Jacobs. Lambertian reflectance and linear subspaces. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 25(2):218–233, 2003. 29, 40

[9] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009. 107

[10] S. R. Becker, E. J. Candès, and M. C. Grant. Templates for convex cone problems with applications to sparse signal recovery. *Mathematical Programming Computation*, 3(3): 165–218, 2011. 217

[11] S. R. Becker, E. J. Candès, and M. C. Grant. TFOCS: Tfocs: Templates for first-order conic solvers. `http://cvxr.com/tfocs/`, Sept. 2012. 85

[12] R. Bell and Y. Koren. Lessons from the netflix prize challenge. *ACM SIGKDD Explorations Newsletter*, 9:75–79, 2007. 10, 24

[13] A. Ben-Tal and A. Nemirovski. Robust convex optimization. *Mathematics of Operations Research*, 23(4):769–805, 1998. 36

[14] J. Bennett, S. Lanning, and N. Netflix. The Netflix Prize. In *In KDD Cup and Workshop in conjunction with KDD*, 2007. 4, 75

[15] D. Bertsimas and M. Sim. The price of robustness. *Operations research*, 52(1):35–53, 2004. 36

[16] W. Bin, D. Chao, S. Defeng, and K.-C. Toh. On the moreau-yosida regularization of the vector k-norm related functions. *Preprint*, 2011. 102

[17] 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. 37, 45, 58, 188, 208, 211

[18] P. Bradley and O. Mangasarian. k-plane clustering. *Journal of Global Optimization*, 16 (1):23–32, 2000. 5, 30

[19] A. M. Buchanan and A. W. Fitzgibbon. Damped newton algorithms for matrix factorization with missing data. In *IJCV*, volume 2, pages 316–322, 2005. 74, 75, 82, 83, 86, 89, 90, 118, 119, 217

[20] E. Candès. The restricted isometry property and its implications for compressed sensing. *Comptes Rendus Mathematique*, 346(9):589–592, 2008. 3, 32

[21] E. Candes and Y. Plan. Matrix completion with noise. *Proc. IEEE*, 98(6):925–936, 2010. 75, 79, 85, 114

[22] E. Candes and Y. Plan. Matrix completion with noise. *Proceedings of the IEEE*, 98: 925–936, 2010. 2, 4, 11, 14, 15, 79

[23] E. Candès and Y. Plan. Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements. *IEEE Info. Theory*, 57:2342–2359, 2011. 149, 152, 153

[24] E. Candes and B. Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9(6):717–772, 2009. ISSN 1615-3375. 2, 4, 75, 79

[25] E. Candès and B. Recht. Exact matrix completion via convex optimization. *Foundations of Computational mathematics*, 9(6):717–772, 2009. 191

[26] E. Candes and T. Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE Info. Theory*, 56:2053–2080, 2010. 22, 154

[27] E. Candès, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *Journal of the ACM*, 58(3), May 2011. 2, 4, 75, 76, 79, 107, 191

[28] E. J. Candes and T. Tao. Decoding by linear programming. *Information Theory, IEEE Transactions on*, 51(12):4203–4215, 2005. 2, 3, 4

[29] E. J. Candès and T. Tao. The power of convex relaxation: Near-optimal matrix completion. *Information Theory, IEEE Transactions on*, 56(5):2053–2080, 2010. 79

[30] E. J. Candes, J. K. Romberg, and T. Tao. Stable signal recovery from incomplete and inaccurate measurements. *Communications on pure and applied mathematics*, 59(8): 1207–1223, 2006. 3, 4

[31] E. J. Candes, M. B. Wakin, and S. P. Boyd. Enhancing sparsity by reweighted ℓ_1 minimization. *Journal of Fourier Analysis and Applications*, 14(5-6):877–905, 2008. 108

[32] E. Cand?s and B. Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9(6):717–772, 2009. ISSN 1615-3375. 4, 13, 79

[33] A. Castrodad and G. Sapiro. Sparse modeling of human actions from motion imagery. *International journal of computer vision*, 100(1):1–15, 2012. 4

[34] V. Chandrasekaran, S. Sanghavi, P. Parrilo, and A. Willsky. Rank-sparsity incoherence for matrix decomposition. *SIAM Journal on Optimization*, 21:572–596, 2011. 80

[35] G. Chen and G. Lerman. Spectral curvature clustering (scc). *International Journal of Computer Vision*, 81(3):317–330, 2009. 5, 30

[36] P. Chen. Optimization algorithms on subspaces: Revisiting missing data problem in low-rank matrix. *IJCV*, 80(1):125–142, 2008. ISSN 0920-5691. 75, 83, 87, 90, 92, 95, 217

[37] P. Chen. Optimization algorithms on subspaces: Revisiting missing data problem in low-rank matrix. *IJCV*, 80:125–142, 2008. 12, 13

[38] P. Chen. Hessian matrix vs. gauss-newton hessian matrix. *SIAM Journal on Numerical Analysis*, 49(4):1417–1435, 2011. 89, 95

[39] Y. Chen, A. Jalali, S. Sanghavi, and C. Caramanis. Low-rank matrix recovery from errors and erasures. In *Information Theory Proceedings (ISIT), 2011 IEEE International Symposium on*, pages 2313–2317. IEEE, 2011. 75, 79, 80

[40] Z. Cheng and N. Hurley. Robustness analysis of model-based collaborative filtering systems. In *AICS'09*, pages 3–15, 2010. 23

[41] F. H. Clarke. *Optimization and Nonsmooth Analysis*. John Wiley and Sons, 1983. ISBN 047187504X. 105

[42] J. Costeira and T. Kanade. A multibody factorization method for independently moving objects. *International Journal of Computer Vision*, 29(3):159–179, 1998. 29, 40

[43] J. Costeira and T. Kanade. *A multi-body factorization method for motion analysis*. Springer, 2000. 5

[44] S. Dasgupta and A. Gupta. An elementary proof of a theorem of johnson and lindenstrauss. *Random Structures & Algorithms*, 22(1):60–65, 2002. 176, 205

[45] K. Davidson and S. Szarek. Local operator theory, random matrices and banach spaces. *Handbook of the geometry of Banach spaces*, 1:317–366, 2001. 22, 156, 205

[46] A. Del Bue, J. Xavier, L. Agapito, and M. Paladini. Bilinear modeling via augmented lagrange multipliers (balm). *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(8):1496 –1508, August 2012. 5, 78, 82, 100, 110, 122, 124, 217

[47] D. Donoho. De-noising by soft-thresholding. *Information Theory, IEEE Transactions on*, 41(3):613–627, 1995. 107, 188

[48] D. Donoho. High-dimensional data analysis: The curses and blessings of dimensionality. aide-memoire of a lecture at ams conference on math challenges of 21st century, 2000. 1, 2

[49] D. Donoho, M. Elad, and V. Temlyakov. Stable recovery of sparse overcomplete representations in the presence of noise. *Information Theory, IEEE Transactions on*, 52(1): 6–18, 2006. 2, 32

[50] D. L. Donoho. Compressed sensing. *Information Theory, IEEE Transactions on*, 52(4): 1289–1306, 2006. 3

[51] P. Drineas, I. Kerenidis, and P. Raghavan. Competitive recommendation systems. In *STOC*, pages 82–90, 2002. 10

[52] M. Elad and M. Aharon. Image denoising via learned dictionaries and sparse representation. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 1, pages 895–900. IEEE, 2006. 4

[53] E. Elhamifar and R. Vidal. Sparse subspace clustering. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 2790–2797. IEEE, 2009. 4, 5, 30, 48

[54] E. Elhamifar and R. Vidal. Clustering disjoint subspaces via sparse representation. In *ICASSP'11*, pages 1926–1929. IEEE, 2010. 30

[55] E. Elhamifar and R. Vidal. Sparse manifold clustering and embedding. *Advances in Neural Information Processing Systems*, 24:55–63, 2011. 132

[56] E. Elhamifar and R. Vidal. Sparse subspace clustering: Algorithm, theory, and applications. *to appear in IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2012. 5, 7, 32, 33, 48

[57] E. Elhamifar and R. Vidal. Sparse subspace clustering: Algorithm, theory, and applications. *to appear in IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2013. ix, 53, 68, 69, 189

[58] A. Eriksson and A. Van Den Hengel. Efficient computation of robust low-rank matrix approximations in the presence of missing data using the l1 norm. *CVPR*, pages 771–778, 2010. ISSN 1424469848. 75, 78, 84, 109, 110, 113, 117, 217

[59] B. Eriksson, L. Balzano, and R. Nowak. High rank matrix completion. In *AI Stats'12*, 2012. 5, 29, 40, 46, 131

[60] S. Friedland, A. Niknejad, M. Kaveh, and H. Zare. An Algorithm for Missing Value Estimation for DNA Microarray Data. In *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, volume 2, page II. IEEE, May 2006. ISBN 1-4244-0469-X. doi: 10.1109/ICASSP.2006.1660537. 4, 74

[61] S. Funk. Netflix update: Try this at home. `http://sifter.org/~simon/journal/20061211.html`, 2006. 74, 83

[62] K. R. Gabriel and S. Zamir. Lower rank approximation of matrices by least squares with any choice of weights. *Technometrics*, 21(4):489–498, 1979. 81

[63] A. Ganesh, J. Wright, X. Li, E. J. Candes, and Y. Ma. Dense error correction for low-rank matrices via principal component pursuit. In *Information Theory Proceedings (ISIT), 2010 IEEE International Symposium on*, pages 1513–1517. IEEE, 2010. 80

[64] Z. Gao, L.-F. Cheong, and M. Shan. Block-sparse rpca for consistent foreground detection. In *Computer Vision–ECCV 2012*, pages 690–703. Springer, 2012. 4

[65] M. Grant and S. Boyd. Graph implementations for nonsmooth convex programs. In V. Blondel, S. Boyd, and H. Kimura, editors, *Recent Advances in Learning and Control*, Lecture Notes in Control and Information Sciences, pages 95–110. Springer-Verlag Limited, 2008. 217

[66] M. Grant and S. Boyd. CVX: Matlab software for disciplined convex programming, version 2.0 beta. `http://cvxr.com/cvx`, Sept. 2012. 85

[67] P. Gritzmann and V. Klee. Computational complexity of inner and outerj-radii of polytopes in finite-dimensional normed spaces. *Mathematical programming*, 59(1):163–213, 1993. 54

[68] R. Hartley and F. Schaffalitzky. Powerfactorization: 3d reconstruction with missing or uncertain data. In *Australia-Japan advanced workshop on computer vision*, volume 74, pages 76–85, 2003. 4, 5, 74, 82

[69] R. Hartley and A. Zisserman. *Multiple view geometry in computer vision*, volume 2. Cambridge Univ Press, 2000. 67

[70] T. Hastie and P. Simard. Metrics and models for handwritten character recognition. *Statistical Science*, pages 54–65, 1998. 40

[71] J. He, L. Balzano, and J. Lui. Online robust subspace tracking from partial information. *arXiv preprint arXiv:1109.3827*, 2011. 78, 84, 109, 126, 217

[72] B. Honigman. 100 fascinating social media statistics and figures from 2012, huffington post. `http://www.huffingtonpost.com/brian-honigman/100-fascinating-social-me_b_2185281.html`, Nov. 2012. 1

[73] B. K. Horn. Height and gradient from shading. *International journal of computer vision*, 5(1):37–75, 1990. 124

[74] K. Huang and S. Aviyente. Sparse representation for signal classification. *Advances in neural information processing systems*, 19:609, 2007. 4

[75] N. Hurley and S. Rickard. Comparing measures of sparsity. *Information Theory, IEEE Transactions on*, 55(10):4723–4741, 2009. 59

[76] P. Jain, P. Netrapalli, and S. Sanghavi. Low-rank matrix completion using alternating minimization. *arXiv preprint arXiv:1212.0467*, 2012. 5, 13, 76, 85, 131

[77] A. Jalali, Y. Chen, S. Sanghavi, and H. Xu. Clustering partially observed graphs via convex optimization. In *ICML'11*, pages 1001–1008. ACM, 2011. 5, 29, 40

[78] I. Jolliffe. *Principal component analysis*, volume 487. Springer-Verlag New York, 1986. 2

[79] K. Kanatani. Motion segmentation by subspace separation and model selection. In *ICCV'01*, volume 2, pages 586–591. IEEE, 2001. 30

[80] Q. Ke and T. Kanade. Robust l1 norm factorization in the presence of outliers and missing data by alternative convex programming. In *CVPR*, volume 1, pages 739–746, 2005. 83, 109

[81] R. Keshavan, A. Montanari, and S. Oh. Low-rank matrix completion with noisy observations: a quantitative comparison. In *Communication, Control, and Computing*, pages 1216–1222, 2009. 81

[82] R. Keshavan, A. Montanari, and S. Oh. Matrix completion from noisy entries. *JMLR*, 11:2057–2078, 2010. 11, 14

[83] R. Keshavan, A. Montanari, and S. Oh. Matrix completion from a few entries. *IEEE Info. Theory*, 56:2980–2998, 2010. 11

[84] F. Kiraly and R. Tomioka. A combinatorial algebraic approach for the identifiability of low-rank matrix completion. In J. Langford and J. Pineau, editors, *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, ICML '12, pages 967–974, New York, NY, USA, July 2012. Omnipress. ISBN 978-1-4503-1285-1. 5, 84

[85] K. Koh, S.-J. Kim, and S. Boyd. An interior-point method for large-scale l1-regularized logistic regression. *Journal of Machine learning research*, 8(8):1519–1555, 2007. 4

[86] Y. Koren. The bellkor solution to the netflix grand prize. *Netflix prize documentation*, 2009. 81

[87] Y. Koren, R. Bell, and C. Volinsky. Matrix factorization techniques for recommender systems. *IEEE Tran. Computer*, 42:30–37, 2009. 5, 9, 13, 74, 75, 76, 81

[88] P. A. Lachenbruch and M. Goldstein. Discriminant analysis. *Biometrics*, pages 69–85, 1979. 2

[89] S. Lam and J. Riedl. Shilling recommender systems for fun and profit. In *WWW'04*, pages 393–402, 2004. 23

[90] F. Lauer and C. Schnorr. Spectral clustering of linear subspaces for motion segmentation. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 678–685. IEEE, 2009. 60

[91] K.-C. Lee, J. Ho, and D. J. Kriegman. Acquiring linear subspaces for face recognition under variable lighting. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 27(5):684–698, 2005. 109

[92] D. Legland. geom3d toolbox [computer software], 2009. URL `http://www.mathworks.com/matlabcentral/fileexchange/24484-geom3d`. 185

[93] X. Li. Compressed sensing and matrix completion with constant proportion of corruptions. *Constructive Approximation*, 37(1):73–99, 2013. 79, 80

[94] Z. Lin, R. Liu, and Z. Su. Linearized alternating direction method with adaptive penalty for low-rank representation. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 612–620, 2011. 58, 210

[95] A. Litvak, A. Pajor, M. Rudelson, and N. Tomczak-Jaegermann. Smallest singular value of random matrices and geometry of random polytopes. *Advances in Mathematics*, 195: 491–523, 2005. 22

[96] G. Liu, Z. Lin, and Y. Yu. Robust subspace segmentation by low-rank representation. In *International Conference on Machine Learning*, volume 3, 2010. 6, 30

[97] G. Liu, H. Xu, and S. Yan. Exact subspace segmentation and outlier detection by low-rank representation. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2012. 48, 54

[98] G. Liu, Z. Lin, S. Yan, J. Sun, Y. Yu, and Y. Ma. Robust recovery of subspace structures by low-rank representation. *IEEE Trans. on Pattern Analysis and Machine Intelligence (TPAMI)*, 2013. 6, 48, 54, 57

[99] Y. Liu, D. Sun, and K. Toh. An implementable proximal point algorithmic framework for nuclear norm minimization. *Mathematical Programming*, 133(1-2):1–38, 2009. ISSN 0025-5610. 102

[100] P. Loh and M. Wainwright. High-dimensional regression with noisy and missing data: Provable guarantees with nonconvexity. *The Annals of Statistics*, 40(3):1637–1664, 2012. 31

[101] K. Mitra, S. Sheorey, and R. Chellappa. Large-scale matrix factorization with missing data under additional constraints. *NIPS*, 23:1642–1650, 2010. 11, 13

[102] B. Mobasher, R. Burke, and J. Sandvig. Model-based collaborative filtering as a defense against profile injection attacks. In *AAAI'06*, volume 21, page 1388, 2006. 23

[103] B. Mobasher, R. Burke, R. Bhaumik, and C. Williams. Toward trustworthy recommender systems: An analysis of attack models and algorithm robustness. *ACM Tran. Inf. Tech.*, 7:23, 2007. 23

[104] B. Nasihatkon and R. Hartley. Graph connectivity in sparse subspace clustering. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 2137–2144. IEEE, 2011. 6, 46, 48, 50, 56

[105] S. Negahban and M. J. Wainwright. Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. *The Journal of Machine Learning Research*, 13: 1665–1697, 2012. 79

[106] A. Ng, M. Jordan, Y. Weiss, et al. On spectral clustering: Analysis and an algorithm. In *NIPS'02*, volume 2, pages 849–856, 2002. 32, 49

[107] S. Oh, A. Montanari, and A. Karbasi. Sensor network localization from local connectivity: Performance analysis for the mds-map algorithm. In *Information Theory Workshop (ITW), 2010 IEEE*, pages 1–5. IEEE, 2010. 74

[108] T. Okatani and K. Deguchi. On the wiberg algorithm for matrix factorization in the presence of missing components. *IJCV*, 72(3):329–337, 2007. ISSN 0920-5691. 75, 83, 87, 90, 217

[109] T. Okatani and K. Deguchi. On the wiberg algorithm for matrix factorization in the presence of missing components. *IJCV*, 72:329–337, 2007. 13

[110] M. L. Overton. NLCG: Nonlinear conjugate gradient. `http://www.cs.nyu.edu/faculty/overton/software/nlcg/index.html`, n.d. 217

[111] M. Paladini, A. D. Bue, M. Stosic, M. Dodig, J. Xavier, and L. Agapito. Factorization for non-rigid and articulated structure using metric projections. *CVPR*, pages 2898–2905, 2009. doi: http://doi.ieeecomputersociety.org/10.1109/CVPRW.2009.5206602. 5, 74, 76, 81

[112] Y. Peng, A. Ganesh, J. Wright, W. Xu, and Y. Ma. Rasl: Robust alignment by sparse and low-rank decomposition for linearly correlated images. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 763–770. IEEE, 2010. 4

[113] V. Rabaud. Vincent's Structure from Motion Toolbox. `http://vision.ucsd.edu/~vrabaud/toolbox/`, n.d. 217

[114] B. Recht. A simpler approach to matrix completion. *arXiv preprint arXiv:0910.0651*, 2009. 4, 79

[115] E. Richard, P. Savalle, and N. Vayatis. Estimation of simultaneously sparse and low rank matrices. In *Proc. International Conference on Machine learning (ICML'12)*, 2012. 49

[116] M. Rudelson and R. Vershynin. Smallest singular value of a random rectangular matrix. *Communications on Pure and Applied Mathematics*, 62:1707–1739, 2009. 22, 156, 205

[117] W. Rudin. *Real and complex analysis*. Tata McGraw-Hill Education, 1987. 102

[118] R. Serfling. Probability inequalities for the sum in sampling without replacement. *The Annals of Statistics*, 2:39–48, 1974. 149

[119] X. Shi and P. S. Yu. Limitations of matrix completion via trace norm minimization. *ACM SIGKDD Explorations Newsletter*, 12(2):16–20, 2011. 4, 76

[120] M. Siegler. Eric schmidt: Every 2 days we create as much information as we did up to 2003, techcrunch. `http://techcrunch.com/2010/08/04/schmidt-data/`, Aug. 2010. 1

[121] J. Silverstein. The smallest eigenvalue of a large dimensional wishart matrix. *The Annals of Probability*, 13:1364–1368, 1985. 22, 156, 205

[122] A. P. Singh and G. J. Gordon. A unified view of matrix factorization models. In *Machine Learning and Knowledge Discovery in Databases*, pages 358–373. Springer, 2008. 5, 10

[123] A. M.-C. So and Y. Ye. Theory of semidefinite programming for sensor network localization. *Mathematical Programming*, 109(2-3):367–384, 2007. 4

[124] M. Soltanolkotabi and E. Candes. A geometric analysis of subspace clustering with outliers. *The Annals of Statistics*, 40(4):2195–2238, 2012. 5, 7, 30, 31, 34, 35, 36, 38,

39, 41, 42, 48, 51, 52, 53, 54, 55, 58, 163, 170, 175, 181, 182, 184, 185, 204, 206, 207, 208

[125] D. A. Spielman, H. Wang, and J. Wright. Exact recovery of sparsely-used dictionaries. *arXiv preprint arXiv:1206.5882*, 2012. 4

[126] N. Srebro. *Learning with matrix factorizations*. PhD thesis, M.I.T., 2004. 4, 10

[127] N. Srebro and T. Jaakkola. Weighted low-rank approximations. In *Proceedings of the 20th International Conference on Machine Learning (ICML-2003)*, volume 20, page 720, 2003. 81, 83, 87, 217

[128] Statistic Brain. Social networking statistics-Statistic Brain. `http://www.statisticbrain.com/social-networking-statistics/`, Nov. 2012. 1

[129] G. Stewart. Perturbation theory for the singular value decomposition. 1998. 159, 213

[130] G. Stewart and J. Sun. *Matrix perturbation theory*. Academic press New York, 1990. 13

[131] P. Sturm and B. Triggs. A factorization based algorithm for multi-image projective structure and motion. In *Computer VisionłECCV'96*, pages 709–720. Springer, 1996. 81

[132] X. Su and T. Khoshgoftaar. A survey of collaborative filtering techniques. *Adv. in AI*, 2009:4, 2009. 9

[133] G. Takács, I. Pilászy, B. Németh, and D. Tikk. Investigation of various matrix factorization methods for large recommender systems. In *ICDMW'08*, pages 553–562, 2008. 13

[134] K. Toh and S. Yun. An accelerated proximal gradient algorithm for nuclear norm regularized linear least squares problems. *Pacific J. Optim*, 6:615–640, 2010. 85, 106, 107

[135] C. Tomasi and T. Kanade. Shape and motion from image streams under orthography: a factorization method. *IJCV*, 9(2):137–154, 1992. ISSN 0920-5691. 5, 76, 81

[136] R. Tron and R. Vidal. A benchmark for the comparison of 3-d motion segmentation algorithms. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8. IEEE, 2007. 6, 30, 48, 67

[137] P. Tseng. Nearest q-flat to m points. *Journal of Optimization Theory and Applications*, 105(1):249–252, 2000. 5

[138] S. Veres. Geometric bounding toolbox 7.3 [computer software], 2006. URL `http://www.mathworks.com/matlabcentral/fileexchange/11678-polyhedron-and-polytope-computations`. 185

[139] R. Vidal. Subspace clustering. *Signal Processing Magazine, IEEE*, 28(2):52–68, 2011. 30

[140] R. Vidal, Y. Ma, and S. Sastry. Generalized principal component analysis (gpca). *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(12):1945–1959, 2005. 5, 30, 48

[141] U. Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416, 2007. 60

[142] Y.-X. Wang and H. Xu. Stability of matrix factorization for collaborative filtering. In J. Langford and J. Pineau, editors, *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, ICML '12, pages 417–424, July 2012. 5, 6, 9, 85

[143] Y.-X. Wang and H. Xu. Noisy sparse subspace clustering. In S. Dasgupta and D. Mcallester, editors, *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, volume 28, pages 89–97. JMLR Workshop and Conference Proceedings, 2013. 6, 7, 29, 48, 57, 59

[144] Y.-X. Wang, C. M. Lee, L.-F. Cheong, and K. C. Toh. Practical matrix completion and corruption recovery using proximal alternating robust subspace minimization. *Under review for publication at IJCV*, 2013. 7, 74

[145] Y.-X. Wang, H. Xu, and C. Leng. Provable subspace clustering: When LRR meets SSC. *To appear at Neural Information Processing Systems (NIPS-13)*, 2013. 7, 47

[146] Z. Wen. Solving a low-rank factorization model for matrix completion by a nonlinear successive over-relaxation algorithm. *Rice University CAAM Technical Report*, pages 1–24, 2010. 11

[147] Z. Wen and W. Yin. A feasible method for optimization with orthogonality constraints. *Mathematical Programming*, pages 1–38, 2013. 83, 87, 217

[148] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma. Robust face recognition via sparse representation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(2):210–227, 2009. 4

[149] L. Wu, A. Ganesh, B. Shi, Y. Matsushita, Y. Wang, and Y. Ma. Robust photometric stereo via low-rank matrix completion and recovery. In *ACCV*, pages 703–717, 2011. ISBN 978-3-642-19317-0. 4, 75, 110, 121, 122

[150] J. Xavier, A. Del Bue, L. Agapito, and M. Paladini. Convergence analysis of balm. *Technical report*, 2011. 100

[151] L. Xiong, X. Chen, and J. Schneider. Direct robust matrix factorization for anomaly detection. *ICDM*, 2011. 84

[152] J. Yang, D. Sun, and K.-C. Toh. A proximal point algorithm for log-determinant optimization with group lasso regularization. *SIAM Journal on Optimization*, 23(2):857–893, 2013. 102

[153] A. Zhang, N. Fawaz, S. Ioannidis, and A. Montanari. Guess who rated this movie: Identifying users through subspace clustering. *arXiv preprint arXiv:1208.1544*, 2012. 29

[154] S. Zhou, G. Aggarwal, R. Chellappa, and D. Jacobs. Appearance characterization of linear lambertian objects, generalized photometric stereo, and illumination-invariant face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(2): 230–245, 2007. 40

[155] Z. Zhou, X. Li, J. Wright, E. Candes, and Y. Ma. Stable principal component pursuit. In *Information Theory Proceedings (ISIT), 2010 IEEE International Symposium on*, pages 1518–1522. IEEE, 2010. 2, 4, 79, 80

Appendices

Appendix A

Appendices for Chapter 2

A.1 Proof of Theorem 2.2: Partial Observation Theorem

In this appendix we prove Theorem 2.2. The proof involves a covering number argument and a concentration inequality for sampling without replacement. The two lemmas are stated below.

Lemma A.1 (Hoeffding Inequality for Sampling without Replacement [118]). *Let $X = [X_1, ..., X_n]$ be a set of samples taken without replacement from a distribution $\{x_1, ...x_N\}$ of mean u and variance σ^2. Denote $a \triangleq \max_i x_i$ and $b \triangleq \min_i x_i$. Then we have:*

$$Pr(|\frac{1}{n}\sum_{i=1}^{n} X_i - u| \geq t) \leq 2\exp(-\frac{2nt^2}{(1-\frac{n-1}{N})(b-a)^2}). \tag{A.1}$$

Lemma A.2 (Covering number for low-rank matrices of bounded size). *Let $S_r = \{X \in \mathbb{R}^{n_1 \times n_2} : rank(X) \leq r, \|X\|_F \leq K\}$. Then there exists an ϵ-net $\bar{S}_r$ for the Frobenius norm obeying*

$$|\bar{S}_r(\epsilon)| \leq (9K/\epsilon)^{(n_1+n_2+1)r}.$$

This Lemma is essentially the same as Lemma 2.3 of [23], with the only difference being the range of $\|X\|_F$: instead of having $\|X\|_F = 1$, we have $\|X\|_F \leq K$. The proof is given in the next section of Appendix.

Proof of Theorem 2.2. Fix $X \in S_r$. Define the following to lighten notations

$$\hat{u}(X) = \frac{1}{|\Omega|}\|P_\Omega(X - \widehat{Y})\|_F^2 = (\hat{\mathcal{L}}(X))^2,$$
$$u(X) = \frac{1}{mn}\|X - \widehat{Y}\|_F^2 = (\mathcal{L}(X))^2.$$

Notice that $\left\{(X_{ij} - \widehat{Y}_{ij})^2\right\}_{ij}$ form a distribution of nm elements, u is its mean, and $\hat{u}$ is the mean of $|\Omega|$ random samples drawn without replacement. Hence, by Lemma A.1:

$$Pr(|\hat{u}(X) - u(X)| > t) \leq 2\exp\left(-\frac{2|\Omega|mnt^2}{(mn - |\Omega| + 1)M^2}\right), \tag{A.2}$$

where $M \triangleq \max_{ij}(X_{ij} - \widehat{Y}_{ij})^2 \leq 4k^2$. Apply union bound over all $X \in \bar{S}_r(\epsilon)$, we have

$$Pr(\sup_{\bar{X}\in\bar{S}_r(\epsilon)} |\hat{u}(\bar{X}) - u(\bar{X})| > t) \leq 2|\bar{S}_r(\epsilon)|\exp\left(-\frac{2|\Omega|mnt^2}{(mn - |\Omega| + 1)M^2}\right)$$

Equivalently, with probability at least $1 - 2\exp(-n)$.

$$\sup_{\bar{X}\in\bar{S}_r(\epsilon)} |\hat{u}(\bar{X}) - u(\bar{X})| \leq \sqrt{\frac{M^2}{2}(n + \log|\bar{S}_r(\epsilon)|)\left(\frac{1}{|\Omega|} - \frac{1}{mn} + \frac{1}{mn|\Omega|}\right)}.$$

Notice that $\|X\|_F \leq \sqrt{mn}k$. Hence substituting Lemma A.2 into the equation, we get:

$$\begin{aligned}&\sup_{\bar{X}\in\bar{S}_r(\epsilon)} |\hat{u}(\bar{X}) - u(\bar{X})| \\ &\leq \left[\frac{M^2}{2}\left(n + (m+n+1)r\log\frac{9k\sqrt{mn}}{\epsilon}\right)\left(\frac{1}{|\Omega|} - \frac{1}{mn} + \frac{1}{mn|\Omega|}\right)\right]^{\frac{1}{2}} := \xi(\Omega)\end{aligned}$$

where we define $\xi(\Omega)$ for convenience. Recall that $\hat{u}(\bar{X}) = (\hat{\mathcal{L}}(X))^2$ and $u(\bar{X}) = (\mathcal{L}(X))^2$. Notice that for any non-negative a and b, $a^2 + b^2 \leq (a+b)^2$. Hence the

following inequalities hold for all $\bar{X} \in \bar{S}_r(\epsilon)$:

$$(\hat{\mathcal{L}}(\bar{X}))^2 \leq (\mathcal{L}(\bar{X}))^2 + \xi(\Omega) \leq (\mathcal{L}(\bar{X}) + \sqrt{\xi(\Omega)})^2,$$
$$(\mathcal{L}(\bar{X}))^2 \leq (\hat{\mathcal{L}}(\bar{X}))^2 + \xi(\Omega) \leq (\hat{\mathcal{L}}(\bar{X}) + \sqrt{\xi(\Omega)})^2,$$

which implies

$$\sup_{\bar{X} \in \bar{S}_r} |\hat{\mathcal{L}}(\bar{X}) - \mathcal{L}(\bar{X})| \leq \sqrt{\xi(\Omega)}.$$

To establish the theorem, we need to relate S_r and $\bar{S}_r(\epsilon)$. For any $X \in S_r$, there exists $c(X) \in \bar{S}_r(\epsilon)$ such that:

$$\|X - c(X)\|_F \leq \epsilon; \qquad \|P_\Omega(X - c(X))\|_F \leq \epsilon;$$

which implies,

$$|\mathcal{L}(X) - \mathcal{L}(c(X))| = \frac{1}{\sqrt{mn}} \left| \|X - \widehat{Y}\|_F - \|c(X) - \widehat{Y}\|_F \right| \leq \frac{\epsilon}{\sqrt{mn}};$$
$$|\hat{\mathcal{L}}(X) - \hat{\mathcal{L}}(c(X))| = \frac{1}{\sqrt{|\Omega|}} \left| \|P_\Omega(X - \widehat{Y})\|_F - \|P_\Omega(c(X) - \widehat{Y})\|_F \right| \leq \frac{\epsilon}{\sqrt{|\Omega|}}.$$

Thus we have,

$$\begin{aligned}
&\sup_{X \in S_r} |\hat{\mathcal{L}}(X) - \mathcal{L}(X)| \\
\leq &\sup_{X \in S_r} \left\{ |\hat{\mathcal{L}}(X) - \hat{\mathcal{L}}(c(X))| + |\mathcal{L}(c(X)) - \mathcal{L}(X)| + |\hat{\mathcal{L}}(c(X)) - \mathcal{L}(c(X))| \right\} \\
\leq &\frac{\epsilon}{\sqrt{|\Omega|}} + \frac{\epsilon}{\sqrt{mn}} + \sup_{X \in S_r} |\hat{\mathcal{L}}(c(X)) - \mathcal{L}(c(X))| \\
\leq &\frac{\epsilon}{\sqrt{|\Omega|}} + \frac{\epsilon}{\sqrt{mn}} + \sup_{\bar{X} \in \bar{S}_r(\epsilon)} |\hat{\mathcal{L}}(\bar{X}) - \mathcal{L}(\bar{X})| \leq \frac{\epsilon}{\sqrt{|\Omega|}} + \frac{\epsilon}{\sqrt{mn}} + \sqrt{\xi(\Omega)}.
\end{aligned}$$

Substitute in the expression of $\xi(\Omega)$ and take $\epsilon = 9k$, we have,

$$\begin{aligned}\sup_{X\in S_r} |\hat{\mathcal{L}}(X) - \mathcal{L}(X)| \leq & 2\frac{\epsilon}{\sqrt{|\Omega|}} + \left(\frac{M^2}{2}\frac{2nr\log(9kn/\epsilon)}{|\Omega|}\right)^{\frac{1}{4}} \\ \leq & \frac{18k}{\sqrt{|\Omega|}} + \sqrt{2}k\left(\frac{nr\log(n)}{|\Omega|}\right)^{\frac{1}{4}} \leq Ck\left(\frac{nr\log(n)}{|\Omega|}\right)^{\frac{1}{4}},\end{aligned}$$

for some universal constant C. This complete the proof. □

A.2 Proof of Lemma A.2: Covering number of low rank matrices

In this appendix, we prove the covering number lemma used in Appendix A.1. As explained in the main text of this thesis, this is an extension of Lemma 2.1 in [23].

Proof of Lemma A.2. This is a two-step proof. First we prove for $\|X\|_F \leq 1$, then we scale it to $\|X\|_F \leq K$.

Step 1: The first part is almost identical to that in Page 14-15 of [23]. We prove via SVD and bound the $\epsilon/3$-covering number of U, Σ and V individually. U and V are bounded the same way. So we only cover the part for of $r \times r$ diagonal singular value matrix Σ.

Now $\|\Sigma\| \leq 1$ instead of $\|\Sigma\| = 1$. $diag(\Sigma)$ lying inside a unit r-sphere (denoted by A). We want to cover this r-sphere with smaller r-sphere of radius $\epsilon/3$ (denoted by B). Then there is a lower bound and an upper bound of the $(\epsilon/3)$-covering number $N(A, B)$.

$$\begin{aligned}\frac{vol(A)}{vol(B)} \leq N(A,B) \leq \bar{N}(A,B) = \bar{N}(A, \frac{B}{2} - \frac{B}{2}) \\ \leq M(A, \frac{B}{2}) \leq \frac{vol(A + B/2)}{vol(B/2)}\end{aligned}$$

where $\bar{N}(A, B)$ is the covering number from inside, and $M(A, B)$ is the number of separated points. Set $B = \frac{B}{2} - \frac{B}{2}$ because B is symmetrical (an n-sphere).

$$(\frac{1}{\epsilon/3})^r \leq N(A,B) \leq (\frac{1+\epsilon/6}{\epsilon/6})^r$$

We are only interested in the upper bound of covering number:

$$N(A,B) \leq (1+6/\epsilon)^r \leq (6/\epsilon + \frac{1}{\epsilon/3}) = (9/\epsilon)^r$$

The inequality is due to the fact that $\epsilon/3 < 1$ (otherwise covering set $B > A$). In fact, we may further tighten the bound by using the fact that all singular values are positive, then A is further constrained in side the first orthant. This should reduce the covering number to its $\frac{1}{2^r}$.

Everything else follows exactly the same way as in [23, Page 14-15].

Step 2: By definition, if $\|X\|_F = 1$, then a finite set of $(9/\epsilon)^{(n_1+n_2+1)r}$ elements are sufficient to ensure that, for every $X \in S_r$, it exists an $\bar{X} \in \bar{S}_r$, such that

$$\|\bar{X} - X\|_F \leq \epsilon$$

Scale both side by K, we get:

$$\|K\bar{X} - KX\|_F \leq K\epsilon$$

let $\beta = K\epsilon$, then the β-net covering number of the set of $\|X\|_F = K$ is:

$$|\bar{S}_r| \leq (9/\epsilon)^{(n_1+n_2+1)r} = (9K/\beta)^{(n_1+n_2+1)r}$$

Revert the notation back to ϵ, the proof is complete. □

A.3 Proof of Proposition 2.1: σ_{min} bound

In this appendix, we develop proof for Proposition 2.1. As is explained in main text of the thesis, σ_{min} can be arbitrarily small in general[1], unless we make assumptions about the structure of matrix. That is why we need strong incoherence property[26] for the proof of Proposition 2.1, which is stated below.

Strong incoherence property with parameter μ, implies that exist $\mu_1, \mu_2 \leq \mu$, such that:

A1 There exists $\mu_1 > 0$ such that for all pair of standard basis vector e_i and e_j (overloaded in both column space and row space of different dimension), there is:

$$\left|\langle e_i, P_U e_j\rangle - \frac{r}{m}1_{i=j}\right| \leq \mu_1\frac{\sqrt{r}}{m}; \qquad \left|\langle e_i, P_V e_j\rangle - \frac{r}{n}1_{i=j}\right| \leq \mu_1\frac{\sqrt{r}}{n}$$

A2 There exists $\mu_2 > 0$ such that for all i, j, the "sign matrix" E defined by $E = UV^T$ satisfies: $|E_{i,j}| = \mu_2\frac{\sqrt{r}}{\sqrt{mn}}$

To interpret A1, again let singular subspace U be denoted by a orthonormal basis matrix N, $P_U = NN^T$. If $i = j$, we have

$$\frac{r - \mu\sqrt{r}}{m} \leq \|n_i\|^2 = \|n_j\|^2 \leq \frac{r + \mu\sqrt{r}}{m}. \tag{A.3}$$

When $i \neq j$, we have $-\frac{\mu\sqrt{r}}{m} \leq n_i^T n_j \leq \frac{\mu\sqrt{r}}{m}$.

Proof of Proposition 2.1. Instead of showing smallest singular value of N_1 directly, we find the $\sigma_{max}(N_2)$ or $\|N_2\|$, and then use the fact that all $\sigma_{min}(N) = 1$ to bound $\sigma_{min}(N_1)$ with their difference.

Let N_2 be of dimension $k \times r$. $\|N_2\| = \|N_2^T\|$, so the maximum singular value equals to $\max_u \|N_2^T u\|$ with u being a unit vector of dimension k. We may consider k

[1]Consider a matrix N with first r rows identity matrix and the rest zero(verify that this is an orthonormal basis matrix). If no observations are taken from first r-rows of user y then all singular values of the N_1 will be zero and (2.4) is degenerate.

a coefficient with $k = [c_1, c_2, ...c_k]^T$. It is easy to see that $c_1^2 + ... + c_k^2 = 1$.

$$\begin{aligned}
\|N_2^T u\|^2 =& u^T N_2 N_2^T u = (c_1 n_1^T + c_2 n_2^T + ... + c_k n_k^T)(c_1 n_1 + c_2 n_2 + ... + c_k n_k) \\
=& (c_1^2 n_1^T n_1 + ... + c_k^2 n_k^T n_k) + 2 \sum_{i<j} c_i c_j n_i^T n_j^T \\
\leq & \sum_{i=1,...,k} (c_i^2) \max_i \|n_i\|^2 + \sum_{i<j} 2|c_i c_j| \max_{i,j} n_i^T n_j \\
\leq & \max_i \|n_i\|^2 + \sum_{i<j} (c_i^2 + c_j^2) \max_{i,j} n_i^T n_j \\
= & \max_i \|n_i\|^2 + (k-1) \sum_{i=1,...,k} (c_i^2) \max_{i,j} n_i^T n_j \\
= & \max_i \|n_i\|^2 + (k-1) \max_{i,j} n_i^T n_j \\
\leq & \frac{r + \mu_1 \sqrt{r}}{m} + (k-1) \frac{\mu_1 \sqrt{r}}{m} = \frac{r}{m} + k \frac{\mu_1 \sqrt{r}}{m}
\end{aligned}$$

The second inequality is by $a^2 + b^2 \geq 2ab$ and last inequality is by the strong incoherence condition.

Similarly, using the $\min_i \|n_i\|^2$ and $\min_{i,j} \|n_i\|^2$ we have a lower bound of $\|N_2^T u\|^2 \geq \frac{r}{m} - k \frac{\mu_1 \sqrt{r}}{m}$. But his bound is not useful/trivial because it decreases with the increase of k, which counters the intuition.

Now, we may express the bound of max singular value $sigma$ in terms of sample rate p of N_1 (hence sample rate of N_2 is $(1-p)$)

$$\sigma_{max}(N_2) \leq \left(\frac{r}{m} + (1-p) \mu_1 \sqrt{r} \right)^{\frac{1}{2}}$$

The desired bound on minimum singular value of N_1 is hence $1 - \sigma_{max}(N_2) = 1 - \left(\frac{r}{m} + (1-p) \mu_1 \sqrt{r} \right)^{\frac{1}{2}}$.

□

A.4 Proof of Proposition 2.2: σ_{min} bound for random matrix

Proof. Without loss of generality we can assume $k > r$ (otherwise the theorem holds trivially), and normalize G such that $\mathbf{E}\|G\|_F^2 = r$. Indeed, random matrix theory [e.g., 45, 116, 121] asserts that G is close to an orthonomal matrix, as the following lemma, adapted from Theorem II.13 of Davidson and Szarek [45], shows:

Lemma A.3. *With probability of at least* $1 - 2\gamma$*,*

$$1 - \sqrt{\frac{r}{m}} - \sqrt{\frac{2\log(1/\gamma)}{m}} \leq \sigma_{min}(G) \leq \sigma_{max}(G) \leq 1 + \sqrt{\frac{r}{m}} + \sqrt{\frac{2\log(1/\gamma)}{m}}.$$

Now let $G = \begin{pmatrix} G_1 \\ G_2 \end{pmatrix}$ such that G_1 is of dimension $k \times r$. Notice that by Lemma A.3, we conclude that there exists an absolute constant such that with probability $1 - C'm^{-10}$,

$$\|G_1 - N_1\| \leq \|G - N\| \leq \sqrt{\frac{r}{m}} + C'\sqrt{\frac{\log m}{m}}.$$

To see this, take compact SVD of $G = USV^T$, U is $m \times r$, S and V are both $r \times r$. In particular, U is orthonormal and V is a rotation matrix. Let $N = UV^T$, then N is an orthonormal basis of G. Furthermore, $G - N = USV^T - UV^T = U(S - I_{r\times r})V^T$ implies $\|G - N\| = |\sigma_{max}(G) - 1|$.

Then using the fact that G_1 is again Gaussian random matrix, we apply Lemma A.3 on G_1 to obtain

$$Pr\left(\sigma_{min}(G_1) \leq \sqrt{\frac{k}{m}} - \sqrt{\frac{r}{m}} - C'\sqrt{\frac{\log m}{m}}\right) \leq C'm^{-10}.$$

This implies that with probability $1 - 2C'm^{-10}$

$$\sigma_{min}(N_1) \geq \sigma_{min}(G_1) - \|G_1 - N_1\| \geq \sqrt{\frac{k}{m}} - 2\sqrt{\frac{r}{m}} - 2C'\sqrt{\frac{\log m}{m}}.$$

□

A.5 Proof of Proposition 2.4: Weak Robustness for Mass Attack

Proof of Proposition 2.4. First observe $\|(E^{gnd^\perp})\|_F \leq k\sqrt{mn_e}$. Note by assumption, sample rate in E block is capped at $3p/2$, thus $\|P_\Omega(E^{gnd^\perp})\|_F \leq k\sqrt{\frac{3p}{2}mn_e}$. Apply Theorem 2.1, we obtained Frobenious norm error:

$$\begin{aligned}\|\Delta\|_F &\leq \frac{1}{\sqrt{p}}\|P_\Omega(E^{gnd^\perp})\|_F + \|(E^{gnd^\perp})\|_F + |\tau(\Omega)| \\ &= (\sqrt{\frac{3}{2}}+1)k\sqrt{mn_e} + Ck\sqrt{m(n+n_e)}\left(\frac{nr\log(n)}{|\Omega|}\right)^{\frac{1}{4}}.\end{aligned}$$

Simplify the equation by absorbing small terms into constant, we get:

$$\|\Delta\|_F \leq Ck\left[\sqrt{mn_e} + \left(\frac{n^3 r\log(n)}{p}\right)^{\frac{1}{4}}\right]$$

By Theorem 2.3:

$$\|\mathbb{P}^{gnd} - \mathbb{P}^{N^*}\| \leq \sqrt{2}\frac{\|\Delta\|}{\delta} = \rho.$$

and δ is greater than $\sigma_r - \sigma_1(E^{gnd^\perp})$.

With condition number κ:

$$\begin{aligned}\sigma_r = \frac{\|Y\|_2}{\kappa} &\geq \frac{\|Y\|_F}{\kappa\sqrt{r}} = \frac{\sqrt{mn\mathbf{E}|Y_{i,j}|^2}}{\kappa\sqrt{r}} \qquad \text{(A.4)} \\ &\geq \frac{\sqrt{mn\mathbf{E}|Y_{i,j}|^2}\sigma_1(E^{gnd^\perp})}{\kappa\sqrt{r}\|E^{gnd^\perp}\|_F} \geq \frac{\sqrt{n\mathbf{E}|Y_{i,j}|^2}\sigma_1(E^{gnd^\perp})}{k\kappa\sqrt{r}\sqrt{n_e}} \\ &\geq \sigma_1(E^{gnd^\perp})\sqrt{\frac{n\mathbf{E}|Y_{i,j}|^2}{k^2\kappa^2 r}/n_e}. \qquad \text{(A.5)}\end{aligned}$$

Substitute n_e into (A.5), we get $\sigma_r \geq n^{1/4}\sigma_1(E^{gnd^\perp})$, or rather $\sigma_1(E^{gnd^\perp}) \leq$

$\sigma_r/n^{1/4}$. Together with (A.4),

$$\delta \geq (1-1/n^{1/4})\sigma_r = \frac{(1-1/n^{1/4})\sqrt{mn\mathbf{E}|Y_{i,j}|^2}}{\kappa\sqrt{r}}$$

It follows that

$$\begin{aligned}\rho = \frac{\|\Delta\|_F}{\delta} &\leq \frac{Ck\left[\sqrt{mn_e} + \left(\frac{n^3 r\log(n)}{p}\right)^{\frac{1}{4}}\right]\cdot\kappa\sqrt{r}}{(1-1/n^{1/4})\sqrt{mn\mathbf{E}|Y_{i,j}|^2}} \\ &\leq C\left[\frac{1}{n^{1/4}} + \frac{k\kappa}{\sqrt{\mathbf{E}|Y_{i,j}|^2}}\left(\frac{r^3\log(n)}{pn}\right)^{\frac{1}{4}}\right] \qquad \text{(A.6)} \\ &\leq \frac{Ck\kappa}{\sqrt{\mathbf{E}|Y_{i,j}|^2}}\left(\frac{r^3\log(n)}{pn}\right)^{1/4}. \qquad \text{(A.7)}\end{aligned}$$

To reach (A.6), we substitute n_e with its maximum value, which cancels out the $\mathbf{E}|Y_{i,j}|^2$, κ, $\sqrt{r}$ in δ and k as well. $(1-1/n^{1/4})$ is absorbed into the constant C. In the second term in the square brackets, $n^{3/4}$ is canceled out by $(mn)^{1/2}$ with the ratio $\sqrt{n/m}$ absorbed into constant term. Also note that $\frac{k}{\sqrt{\mathbf{E}|Y_{i,j}|^2}} > 1$, $\kappa > 1$, $\frac{r^3\log(n)}{p} > 1$, so the second term is larger than $\frac{1}{n^{1/4}}$ and we may reach (A.7).

Apply Theorem 2.4:

$$\|y^* - y^{gnd}\| \leq \frac{2Ck\kappa\|y\|}{\sigma_{min}\sqrt{\mathbf{E}|Y_{i,j}|^2}}\left(\frac{r^3\log(n)}{pn}\right)^{1/4}, \qquad \text{(A.8)}$$

$$\|e^* - e^{gnd}\| \leq \frac{2Ck\kappa\|e^{gnd^\perp}\|}{\sigma_{min}\sqrt{\mathbf{E}|Y_{i,j}|^2}}\left(\frac{r^3\log(n)}{pn}\right)^{1/4} + \frac{\|e^{gnd^\perp}\|}{\sigma_{min}} = \frac{C\|e^{gnd^\perp}\|}{\sigma_{min}}. \qquad \text{(A.9)}$$

Now let us deal with σ_{min}. By assumption, all user have sample rate of at least $\frac{p}{2}$. By Proposition 2.2 and union bound, we confirm that for some constant c, with probability greater than $1-cn^{-10}$, $\sigma_{min} \geq \sqrt{\frac{p}{2}}$ (relaxed by another $\sqrt{2}$ to get rid of the small terms) for all users.

Summing (A.8) over all users, we get:

$$\begin{aligned}
&\|Y^* - Y\|_F \\
=&\sqrt{\sum_{allusers} \|y^* - y^{gnd}\|^2} = \frac{2Ck\kappa}{\sigma_{min}\sqrt{\mathbf{E}|Y_{i,j}|^2}} \left(\frac{r^3 \log(n)}{pn}\right)^{1/4} \sqrt{\sum_{allusers} \|y\|^2} \\
\leq&\frac{2\sqrt{2}Ck\kappa}{\sqrt{p\mathbf{E}|Y_{i,j}|^2}} \left(\frac{r^3 \log(n)}{pn}\right)^{1/4} \sqrt{mn\mathbf{E}|Y_{i,j}|^2} \leq C_1 \kappa k \sqrt{mn} \left(\frac{r^3 \log(n)}{p^3 n}\right)^{1/4},
\end{aligned}$$

so $RMSE_Y \leq C_1 \kappa k \left(\frac{r^3 \log(n)}{p^3 n}\right)^{1/4}$ is proved.

Similarly from (A.9), $RMSE_E \leq C\frac{\|E^{gnd^\perp}\|_F}{\sqrt{mn_e}}\sqrt{\frac{2}{p}} \leq \frac{C_2 k}{\sqrt{p}}$.

□

A.6 SVD Perturbation Theory

The following theorems in SVD Perturbation Theory [129] are applied in our proof of the subspace stability bound (Theorem 2.3).

1. Weyl's Theorem gives a perturbation bound for singular values.

Lemma A.4 (Weyl)**.**

$$|\hat{\sigma}_i - \sigma_i| \leq \|E\|_2, i = 1, ..., n.$$

2. Wedin's Theorem provides a perturbation bound for singular subspace. To state the Lemma, we need to re-express the singular value decomposition of Y and $\widehat{Y}$ in block matrix form:

$$Y = \begin{pmatrix} L_1 & L_2 & L_3 \end{pmatrix} \begin{pmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} R_1 \\ R_2 \end{pmatrix} \tag{A.10}$$

$$\widehat{Y} = \begin{pmatrix} \hat{L}_1 & \hat{L}_2 & \hat{L}_3 \end{pmatrix} \begin{pmatrix} \hat{\Sigma}_1 & 0 \\ 0 & \hat{\Sigma}_2 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \hat{R}_1 \\ \hat{R}_2 \end{pmatrix} \tag{A.11}$$

Let Φ denotes the canonical angles between $span(L_1)$ and $span(\hat{L}_1)$; let Θ denotes the canonical angle matrix between $span(R_1)$ and $span(\hat{R}_1)$. Also, define residuals:

$$Z = Y\hat{R}_1^T - \hat{L}_1\hat{\Sigma}_1, \qquad S = Y^T\hat{L}_1 - \hat{R}_1^T\hat{\Sigma}_1.$$

The Wedin's Theorem bounds Φ and Θ together using the Frobenious norm of Z and S.

Lemma A.5 (Wedin)**.** *If there is a $\delta > 0$ such that*

$$\min|\sigma(\hat{\Sigma}_1) - \sigma(\Sigma_2)| \geq \delta, \tag{A.12}$$

$$\min\sigma(\hat{\Sigma}_1) \geq \delta, \tag{A.13}$$

then

$$\sqrt{\|\sin\Phi\|_F^2 + \|\sin\Theta\|_F^2} \leq \frac{\sqrt{\|Z\|_F^2 + \|S\|_F^2}}{\delta} \tag{A.14}$$

Besides Frobenious norm, the same result goes for $\|\cdot\|_2$, the spectral norm of everything.

Lemma A.5(Wedin's Theorem) says that if the two separation conditions on singular value (A.12) and (A.13) are satisfied, we can bound the impact of perturbation on the left and right singular subspace simultaneously.

A.7 Discussion on Box Constraint in (2.1)

The box constraint is introduced due to the proof technique used in Section 2.3. We suspect that a more refined analysis may be possible to remove such a constraint. As for results of other sections, such constraint is not needed. Yet, it does not hurt to impose such constraint to (2.3), which will lead to similar results of subspace stability (though much more tedious in proof). Moreover, notice that for sufficiently large k, the solution will remain unchanged with or without the constraint.

On the other hand, we remark that such the box constraint is most natural for the application in collaborative filtering. Since user ratings are usually bounded in a pre-defined range. In real applications, either such box constraint or regularization will be needed to avoid over fitting to the noisy data. This is true regardless whether formulation (2.1) or (2.3) is used.

A.8 Table of Symbols and Notations

For easy reference of the readers, we compiled the following table.

Table A.1: Table of Symbols and Notations

Y	$m \times n$ ground truth rating matrix.
E	$m \times n$ error matrix, in Section 2.6 dummy user matrix.
$\widehat{Y}$	Noisy observation matrix $\widehat{Y} = Y + E$.
Y^*, U^*, V^*	Optimal solution of (2.1) $Y^* = U^* V^{*T}$
$(\cdot)^*, \hat{(\cdot)}, (\cdot)^{gnd}$	Refer to optimal solution, noisy observation, ground truth.
i, j	Item index and user index
r	Rank of ground truth matrix
Ω	The set of indices (i, j) of observed entries.
$\lvert\Omega\rvert$	Cardinality of set Ω.
P_Ω	The projection defined in (2.2).
k	$[-k, k]$ Valid range of user rating.
Δ	Frobenious norm error $\lVert Y^* - Y \rVert_F$
$\mathcal{N}, \mathcal{N}^\perp$	Denote subspace and complement subspace
$N, N^\perp$	Orthonormal basis matrix of $\mathcal{N}, \mathcal{N}^\perp$
N_i	Shortened N with only observed rows in column i
y_i	Observed subset of column i
$\mathbb{P}^{\mathcal{N}}$	Projection matrix to subspace $\mathcal{N}$
$\mathbb{P}_i$	Projection matrix to shortened subspace $\mathrm{span}(N_i)$
τ	The gap of RMSE residual in the proof of Theorem 2.1.
$\mathcal{L}, \widehat{\mathcal{L}}$	Loss function in Theorem 2.2.
ρ	Bounded value of $\lVert \sin(\Theta) \rVert$ of Theorem 2.3.
δ	The r^{th} singular value of Y^* used in Theorem 2.3.
S_r	The collection of all rank-r $m \times n$ matrices.
μ	Coherence parameter in Proposition 2.1
s_{max}	Sparse parameter in Proposition 2.3
κ	Matrix condition number used in Proposition 2.4
p	Sample rate $\frac{\lvert\Omega\rvert}{m(n+n_e)}$ used in Proposition 2.4
C, c, C_1, C_2, C'	Numerical constants
$\sigma_i, \sigma_{min}, \sigma_{max}$	i^{th}, minimum, maximum singular value.
θ_i	i^{th} canonical angle.
Θ, Φ	Diagonal canonical angle matrix.
$\lvert\cdot\rvert$	Either absolute value or cardinality.
$\lVert\cdot\rVert_2$	2-norm of vector/spectral norm of matrix.
$\lVert\cdot\rVert_F$	Frobenious norm of a matrix.
$\lVert\cdot\rVert$	In Theorem 2.3 means both Frobenious norm and spectral norm, otherwise same as $\lVert\cdot\rVert_2$.

Appendix B

Appendices for Chapter 3

B.1 Proof of Theorem 3.1

Our main deterministic result Theorem 3.1 is proved by duality. We first establish a set of conditions on the optimal dual variable of D_0 corresponding to *all* primal solutions satisfying self-expression property. Then we construct such a dual variable ν, hence certify that the optimal solution of P_0 satisfies the LASSO Subspace Detection Property.

B.1.1 Optimality Condition

Define general convex optimization:

$$\min_{c,e} \|c\|_1 + \frac{\lambda}{2}\|e\|^2 \qquad s.t. \quad x = Ac + e. \tag{B.1}$$

We may state an extension of the Lemma 7.1 in Soltanolkotabi and Candes's SSC Proof.

Lemma B.1. *Consider a vector $y \in \mathbb{R}^d$ and a matrix $A \in \mathbb{R}^{d\times N}$. If there exists triplet (c, e, ν) obeying $y = Ac + e$ and c has support $S \subseteq T$, furthermore the dual certificate vector ν satisfies*

$$A_s^T\nu = sgn(c_S), \qquad \nu = \lambda e,$$
$$\|A_{T\cap S^c}^T\nu\|_\infty \leq 1, \quad \|A_{T^c}^T\nu\|_\infty < 1,$$

then all optimal solution (c^*, e^*) to (B.1) *obey $c^*_{T^c} = 0$.*

Proof. For optimal solution (c^*, e^*), we have:

$$\begin{aligned}
&\|c^*\|_1 + \frac{\lambda}{2}\|e^*\|^2 = \|c^*_S\|_1 + \|c^*_{T\cap S^c}\|_1 + \|c^*_{T^c}\|_1 + \frac{\lambda}{2}\|e^*\|^2 \\
\geq &\|c_S\|_1 + \langle sgn(c_S), c^*_S - c_S\rangle + \|c^*_{T\cap S^c}\|_1 + \|c^*_{T^c}\|_1 + \frac{\lambda}{2}\|e\|^2 + \langle \lambda e, e^* - e\rangle \\
= &\|c_S\|_1 + \langle \nu, A_S(c^*_S - c_S)\rangle + \|c^*_{T\cap S^c}\|_1 + \|c^*_{T^c}\|_1 + \frac{\lambda}{2}\|e\|^2 + \langle \nu, e^* - e\rangle \\
= &\|c_S\|_1 + \frac{\lambda}{2}\|e\|^2 + \|c^*_{T\cap S^c}\|_1 - \langle \nu, A_{T\cap S^c}(c^*_{T\cap S^c})\rangle + \|c^*_{T^c}\|_1 - \langle \nu, A_{T^c}(c^*_{T^c})\rangle
\end{aligned} \tag{B.2}$$

To see $\frac{\lambda}{2}\|e^*\|^2 \geq \frac{\lambda}{2}\|e\|^2 + \langle \lambda e, e^* - e\rangle$, note that right hand side equals to $\lambda\left(-\frac{1}{2}e^T e + (e^*)^T e\right)$, which takes a maximal value of $\frac{\lambda}{2}\|e^*\|^2$ when $e = e^*$. The last equation holds because both (c, e) and (c^*, e^*) are feasible solution, such that $\langle \nu, A(c^* - c)\rangle + \langle \nu, e^* - e\rangle = \langle \nu, Ac^* + e^* - (Ac + e)\rangle = 0$. Also, note that $\|c_S\|_1 + \frac{\lambda}{2}\|e\|^2 = \|c\|_1 + \frac{\lambda}{2}\|e\|^2$.

With the inequality constraints of ν given in the Lemma statement, we know

$$\langle \nu, A_{T\cap S^c}(c^*_{T\cap S^c})\rangle = \langle A^T_{T\cap S^c}\nu, (c^*_{T\cap S^c})\rangle \leq \|A^T_{T\cap S^c}\nu\|_\infty \|c^*_{T\cap S^c}\|_1 \leq \|c^*_{T\cap S^c}\|_1.$$

Substitute into (B.2), we get:

$$\|c^*\|_1 + \frac{\lambda}{2}\|e^*\|^2 \geq \|c\|_1 + \frac{\lambda}{2}\|e\|^2 + (1 - \|A^T_{T^c}\nu\|_\infty)\|c^*_{T^c}\|_1,$$

where $(1 - \|A^T_{T^c}\nu\|_\infty)$ is strictly greater than 0.

Using the fact that (c^*, e^*) is an optimal solution, $\|c^*\|_1 + \frac{\lambda}{2}\|e^*\|^2 \leq \|c\|_1 + \frac{\lambda}{2}\|e\|^2$. Therefore, $\|c^*_{T^c}\|_1 = 0$ and (c, e) is also an optimal solution. This concludes the proof. □

Apply Lemma B.1 (same as the Lemma 3.1 in Section 3.4) with $x = x_i^{(\ell)}$ and $A = X_{-i}$, we know that if we can construct a dual certificate ν such that all conditions are satisfied with respect to a feasible solution (c, e) and c satisfy SEP, then the all optimal solution of (3.6) satisfies SEP, in other word $c_i = \left[0, ..., 0, (c_i^{(\ell)})^T, 0, ..., 0\right]^T$.

By definition of LASSO detection property, we must further ensure $\|c_i^{(\ell)}\|_1 \neq 0$

to avoid the trivial solution that $x_i^{(\ell)} = e^*$. This is a non-convex constraint and hard to impose. To this matter, we note that given sufficiently large λ, $\|c_i^{(\ell)}\|_1 \neq 0$ never occurs.

Our strategy of avoiding this trivial solution is hence showing the existence of a λ such that the dual optimal value is smaller than the trivial optimal value, namely:

$$OptVal(\mathbf{D}_0) = \langle x_i, \nu \rangle - \frac{1}{2\lambda}\|\nu\|^2 < \frac{\lambda}{2}\|x_i^{(\ell)}\|^2. \tag{B.3}$$

B.1.2 Constructing candidate dual vector ν

A natural candidate of the dual solution ν is the dual point corresponding to the optimal solution of the following fictitious optimization program.

$$\mathbf{P}_1: \quad \min_{c_i^{(\ell)}, e_i} \|c_i^{(\ell)}\|_1 + \frac{\lambda}{2}\|e_i\|^2 \quad s.t. \quad y_i^{(\ell)} + z_i = (Y_{-i}^{(\ell)} + Z_{-i}^{(\ell)})c_i^{(\ell)} + e_i \tag{B.4}$$

$$\mathbf{D}_1: \quad \max_{\nu} \langle x_i^{(\ell)}, \nu \rangle - \frac{1}{2\lambda}\nu^T\nu \quad s.t. \quad \|(X_{-i}^{(\ell)})^T\nu\|_\infty \leq 1. \tag{B.5}$$

This optimization is feasible because $y_i^{(\ell)} \in span(Y_{-i}^{(\ell)}) = \mathcal{S}_\ell$ so any $c_i^{(\ell)}$ obeying $y_i^{(\ell)} = Y_{-i}^{(\ell)} c_i^{(\ell)}$ and corresponding $e_i = z_i - Z_{-i}^{(\ell)} c_i^{(\ell)}$ is a pair of feasible solution. Then by strong duality, the dual program is also feasible, which implies that for every optimal solution (c, e) of (B.4) with c supported on S, there exist ν satisfying:

$$\left\{ \begin{array}{c} \|((Y_{-i}^{(\ell)})_{S^c}^T + (Z_{-i}^{(\ell)})_{S^c}^T)\nu\|_\infty \leq 1, \quad \nu = \lambda e, \\ ((Y_{-i}^{(\ell)})_S^T + (Z_{-i}^{(\ell)})_S^T)\nu = sgn(c_S). \end{array} \right\}$$

This construction of ν satisfies all conditions in Lemma B.1 with respect to

$$\begin{cases} c_i = [0, ..., 0, c_i^{(\ell)}, 0, ..., 0] \text{ with } c_i^{(\ell)} = c, \\ e_i = e, \end{cases} \tag{B.6}$$

except

$$\left\|[X_1, ..., X_{\ell-1}, X_{\ell+1}, ..., X_L]^T\nu\right\|_\infty < 1,$$

i.e., we must check for all data point $x \in \mathcal{X} \setminus \mathcal{X}^\ell$,

$$|\langle x, \nu \rangle| < 1. \tag{B.7}$$

Showing the solution of (B.5) ν also satisfies (B.7) gives precisely a dual certificate as required in Lemma B.1, hence implies that the candidate solution (B.6) associated with optimal (c, e) of (B.4) is indeed the optimal solution of (3.6).

B.1.3 Dual separation condition

In this section, we establish the conditions required for (B.7) to hold. The idea is to provide an upper bound of $|\langle x, \nu \rangle|$ then make it smaller than 1.

First, we find it appropriate to project ν to the subspace $\mathcal{S}_\ell$ and its complement subspace then analyze separately. For convenience, denote $\nu_1 := \mathbb{P}_{S_\ell}(\nu)$, $\nu_2 := \mathbb{P}_{\mathcal{S}_\ell^c}(\nu)$. Then

$$\begin{aligned} |\langle x, \nu \rangle| =& |\langle y + z, \nu \rangle| \leq |\langle y, \nu_1 \rangle| + |\langle y, \nu_2 \rangle| + |\langle z, \nu \rangle| \\ \leq& \mu(\mathcal{X}_\ell)\|\nu_1\| + \|y\|\|\nu_2\||\cos(\angle(y, \nu_2))| + \|z\|\|\nu\||\cos(\angle(z, \nu))|. \end{aligned} \tag{B.8}$$

To see the last inequality, check that by Definition 3.3, $|\langle y, \frac{\nu_1}{\|\nu_1\|} \rangle| \leq \mu(\mathcal{X}_\ell)$.

Since we are considering general (possibly adversarial) noise, we will use the relaxation $|\cos(\theta)| \leq 1$ for all cosine terms (a better bound under random noise will be given later). Now all we have to do is to bound $\|\nu_1\|$ and $\|\nu_2\|$ (note $\|\nu\| = \sqrt{\|\nu_1\|^2 + \|\nu_2\|^2} \leq \|\nu_1\| + \|\nu_2\|$).

B.1.3.1 Bounding $\|\nu_1\|$

We first bound $\|\nu_1\|$ by exploiting the feasible region of ν_1 in (B.5).

$$\|(X_{-i}^{(\ell)})^T \nu\|_\infty \leq 1$$

is equivalent to $x_i^T \nu \leq 1$ for every x_i that is the column of $X_{-i}^{(\ell)}$. Decompose the condition into

$$y_i^T \nu_1 + (\mathbb{P}_{\mathbb{S}_\ell} z_i)^T \nu_1 + z_i^T \nu_2 \leq 1.$$

Now we relax each of the term into

$$y_i^T \nu_1 + (\mathbb{P}_{\mathbb{S}_\ell} z_i)^T \nu_1 \leq 1 - z_i^T \nu_2 \leq 1 + \delta \|\nu_2\|. \tag{B.9}$$

The relaxed condition contains the feasible region of ν_1 in (B.5).

It turns out that the geometric interpretation of the relaxed constraints gives a upper bound of $\|\nu_1\|$.

Definition B.1 (polar set)**.** *The polar set $\mathcal{K}^o$ of set $\mathcal{K} \in \mathbb{R}^d$ is defined as*

$$\mathcal{K}^o = \left\{ y \in \mathbb{R}^d : \langle x, y \rangle \leq 1 \textit{ for all } x \in \mathcal{K} \right\}.$$

By the polytope geometry, we have

$$\begin{aligned} &\|(Y_{-i}^{(\ell)} + \mathbb{P}_{\mathbb{S}_\ell}(Z_{-i}^{(\ell)}))^T \nu_1\|_\infty \leq 1 + \delta \|\nu_2\| \\ \Leftrightarrow\ &\nu_1 \in \left[\mathcal{P} \left(\frac{Y_{-i}^{(\ell)} + \mathbb{P}_{\mathbb{S}_\ell}(Z_{-i}^{(\ell)})}{1 + \delta \|\nu_2\|} \right) \right]^o := \mathcal{T}^o. \end{aligned} \tag{B.10}$$

Now we introduce the concept of circumradius.

Definition B.2 (circumradius)**.** *The circumradius of a convex body $\mathcal{P}$, denoted by $R(\mathcal{P})$, is defined as the radius of the smallest Euclidean ball containing $\mathcal{P}$.*

The magnitude $\|\nu_1\|$ is bounded by $R(\mathcal{T}^o)$. Moreover, by the the following lemma we may find the circumradius by analyzing the polar set of $\mathcal{T}^o$ instead. By the property of polar operator, polar of a polar set gives the tightest convex envelope of original set, i.e., $(\mathcal{K}^o)^o = conv(\mathcal{K})$. Since $\mathcal{T} = \text{conv}\left(\pm \frac{Y_{-i}^{(\ell)} + \mathbb{P}_{\mathbb{S}_\ell}(Z_{-i}^{(\ell)})}{1 + \delta \|\nu_2\|} \right)$ is convex in the first place, the polar set of $\mathcal{T}^o$ is essentially $\mathcal{T}$.

Lemma B.2. *For a symmetric convex body $\mathcal{P}$, i.e. $\mathcal{P} = -\mathcal{P}$, inradius of $\mathcal{P}$ and circumradius of polar set of $\mathcal{P}$ satisfy:*

$$r(\mathcal{P})R(\mathcal{P}^o) = 1.$$

Lemma B.3. *Given $X = Y + Z$, denote $\rho := \max_i \|\mathbb{P}_{\mathcal{S}} z_i\|$, furthermore $Y \in \mathcal{S}$ where $\mathcal{S}$ is a linear subspace, then we have:*

$$r(\mathrm{Proj}_{\mathcal{S}}(\mathcal{P}(X))) \geq r(\mathcal{P}(Y)) - \rho$$

Proof. First note that projection to subspace is a linear operator, hence $\mathrm{Proj}_{\mathcal{S}}(\mathcal{P}(X)) = \mathcal{P}(\mathbb{P}_{\mathcal{S}} X)$. Then by definition, the boundary set of $\mathcal{P}(\mathbb{P}_{\mathcal{S}} X)$ is

$$\mathcal{B} := \{y \mid y = \mathbb{P}_{\mathcal{S}} X c; \|c\|_1 = 1\}.$$

Inradius by definition is the largest ball containing in the convex body, hence $r(\mathcal{P}(\mathbb{P}_{\mathcal{S}} X)) = \min_{y \in \mathcal{B}} \|y\|$. Now we provide a lower bound of it:

$$\|y\| \geq \|Yc\| - \|\mathbb{P}_{\mathcal{S}} Zc\| \geq r(\mathcal{P}(Y)) - \sum\nolimits_j \|\mathbb{P}_{\mathcal{S}} z_j\| |c_j| \geq r(\mathcal{P}(Y)) - \rho \|c\|_1.$$

This concludes the proof. □

A bound of $\|\nu_1\|$ follows directly from Lemma B.2 and Lemma B.3:

$$\begin{aligned}\|\nu_1\| \leq & (1 + \delta\|\nu_2\|) R(\mathcal{P}(Y_{-i}^{(\ell)} + \mathbb{P}_{\mathcal{S}_\ell}(Z_{-i}^{(\ell)}))) \\ = & \frac{1 + \delta\|\nu_2\|}{r(\mathcal{P}(Y_{-i}^{(\ell)} + \mathbb{P}_{\mathcal{S}_\ell}(Z_{-i}^{(\ell)}))} = \frac{1 + \delta\|\nu_2\|}{r(\mathrm{Proj}_{\mathcal{S}_\ell}(\mathcal{P}(X_{-i}^{(\ell)})))} \leq \frac{1 + \delta\|\nu_2\|}{r\left(\mathcal{Q}_{-i}^{\ell}\right) - \delta_1}. \end{aligned} \tag{B.11}$$

This bound unfortunately depends $\|\nu_2\|$. This can be extremely loose as in general, ν_2 is not well-constrained (see the illustration in Figure B.2 and B.3). That is why we need to further exploit the fact ν is the optimal solution of (B.5), which provides a reasonable bound of $\|\nu_2\|$.

B.1.3.2 Bounding $\|\nu_2\|$

By optimality condition: $\nu = \lambda e_i = \lambda(x_i - X_{-i}c)$ and $\nu_2 = \lambda \mathbb{P}_{\mathcal{S}_\ell^\perp}(x_i - X_{-i}c) = \lambda \mathbb{P}_{\mathcal{S}_\ell^\perp}(z_i - Z_{-i}c)$ so

$$\begin{aligned}\|\nu_2\| &\le \lambda\left(\|\mathbb{P}_{\mathcal{S}_\ell^\perp} z_i\| + \|\mathbb{P}_{\mathcal{S}_\ell^\perp} Z_{-i}c\|\right) \le \lambda(\|\mathbb{P}_{\mathcal{S}_\ell^\perp} z_i\| + \sum_{j\in S}|c_j|\|\mathbb{P}_{\mathcal{S}_\ell^\perp} z_j\|) \\ &\le \lambda(\|c\|_1 + 1)\delta_2 \le \lambda(\|c\|_1+1)\delta \end{aligned} \tag{B.12}$$

Now we will bound $\|c\|_1$. As c is the optimal solution, $\|c\|_1 \le \|c\|_1 + \frac{\lambda}{2}\|e\|^2 \le \|\tilde{c}\|_1 + \frac{\lambda}{2}\|\tilde{e}\|^2$ for any feasible solution $(\tilde{c}, \tilde{e})$. Let $\tilde{c}$ be the solution of

$$\min_c \|c\|_1 \quad s.t. \qquad y_i^{(\ell)} = Y_{-i}^{(\ell)} c, \tag{B.13}$$

then by strong duality, $\|\tilde{c}\|_1 = \max_\nu \left\{\langle \nu, y_i^{(\ell)}\rangle \mid \|[Y_{-i}^{(\ell)}]^T\nu\|_\infty \le 1\right\}$. By Lemma B.2, optimal dual solution $\tilde{\nu}$ satisfies $\|\tilde{\nu}\| \le \frac{1}{r(\mathcal{Q}_{-i}^\ell)}$. It follows that $\|\tilde{c}\|_1 = \langle \tilde{\nu}, y_i^{(\ell)}\rangle = \|\tilde{\nu}\|\|y_i^{(\ell)}\| \le \frac{1}{r(\mathcal{Q}_{-i}^\ell)}$.

On the other hand, $\tilde{e} = z_i - Z_{-i}^{(\ell)}\tilde{c}$, so $\|\tilde{e}\|^2 \le (\|z_i\| + \sum_j \|z_j\||\tilde{c}_j|)^2 \le (\delta + \|\tilde{c}\|_1\delta)^2$, thus: $\|c\|_1 \le \|\tilde{c}\|_1 + \frac{\lambda}{2}\|\tilde{e}\|^2 \le \frac{1}{r(\mathcal{Q}_{-i}^\ell)} + \frac{\lambda}{2}\delta^2\left[1 + \frac{1}{r(\mathcal{Q}_{-i}^\ell)}\right]^2$. This gives the bound we desired:

$$\begin{aligned}\|\nu_2\| &\le \lambda\left(\frac{1}{r(\mathcal{Q}_{-i}^\ell)} + \frac{\lambda}{2}\delta^2\left[1 + \frac{1}{r(\mathcal{Q}_{-i}^\ell)}\right]^2 + 1\right)\delta \\ &= \lambda\delta\left(\frac{1}{r(\mathcal{Q}_{-i}^\ell)} + 1\right) + \frac{\delta}{2}\left\{\lambda\delta\left(\frac{1}{r(\mathcal{Q}_{-i}^\ell)} + 1\right)\right\}^2.\end{aligned}$$

By choosing λ satisfying

$$\lambda\delta^2 \le \frac{2}{1 + 1/r(\mathcal{Q}_{-i}^\ell)}, \tag{B.14}$$

the bound can be simplified to:

$$\|\nu_2\| \le 2\lambda\delta\left(\frac{1}{r(\mathcal{Q}_{-i}^\ell)} + 1\right) \tag{B.15}$$

B.1.3.3 Conditions for $|\langle x, \nu \rangle| < 1$

Putting together (B.8), (B.11) and (B.15), we have the upper bound of $|\langle x, \nu \rangle|$:

$$\begin{aligned}
&|\langle x, \nu \rangle| \leq (\mu(\mathcal{X}_\ell) + \|\mathbb{P}_{\mathbb{S}_\ell} z\|)\|\nu_1\| + (\|y\| + \|\mathbb{P}_{\mathbb{S}_\ell^\perp} z\|)\|\nu_2\| \\
\leq &\frac{\mu(\mathcal{X}_\ell) + \delta_1}{r\big(\mathcal{Q}_{-i}^\ell\big) - \delta_1} + \left(\frac{(\mu(\mathcal{X}_\ell) + \delta_1)\delta}{r\big(\mathcal{Q}_{-i}^\ell\big) - \delta_1} + 1 + \delta \right) \|\nu_2\| \\
\leq &\frac{\mu(\mathcal{X}_\ell) + \delta_1}{r\big(\mathcal{Q}_{-i}^\ell\big) - \delta_1} + 2\lambda\delta(1+\delta)\left(\frac{1}{r(\mathcal{Q}_{-i}^\ell)} + 1 \right) + \frac{2\lambda\delta^2(\mu(\mathcal{X}_\ell) + \delta_1)}{r\big(\mathcal{Q}_{-i}^\ell\big) - \delta_1}\left(\frac{1}{r(\mathcal{Q}_{-i}^\ell)} + 1 \right)
\end{aligned}$$

For convenience, we further relax the second $r(\mathcal{Q}_{-i}^\ell)$ into $r(\mathcal{Q}_{-i}^\ell) - \delta_1$. The dual separation condition is thus guaranteed with

$$\begin{aligned}
&\frac{\mu(\mathcal{X}_\ell) + \delta_1 + 2\lambda\delta(1+\delta) + 2\lambda\delta^2(\mu(\mathcal{X}_\ell) + \delta_1)}{r\big(\mathcal{Q}_{-i}^\ell\big) - \delta_1} \\
&+2\lambda\delta(1+\delta) + \frac{2\lambda\delta^2(\mu(\mathcal{X}_\ell) + \delta_1)}{r\big(\mathcal{Q}_{-i}^\ell\big)(r\big(\mathcal{Q}_{-i}^\ell\big) - \delta_1)} < 1.
\end{aligned}$$

Denote $\rho := \lambda\delta(1+\delta)$, assume $\delta < r\big(\mathcal{Q}_{-i}^\ell\big)$, $(\mu(\mathcal{X}_\ell) + \delta_1) < 1$ and simplify the form with

$$\frac{2\lambda\delta^2(\mu(\mathcal{X}_\ell) + \delta_1)}{r\big(\mathcal{Q}_{-i}^\ell\big) - \delta_1} + \frac{2\lambda\delta^2(\mu(\mathcal{X}_\ell) + \delta_1)}{r\big(\mathcal{Q}_{-i}^\ell\big)(r\big(\mathcal{Q}_{-i}^\ell\big) - \delta_1)} < \frac{2\rho}{r\big(\mathcal{Q}_{-i}^\ell\big) - \delta_1},$$

we get a sufficient condition

$$\mu(\mathcal{X}_\ell) + 3\rho + \delta_1 < (1 - 2\rho)\,(r(\mathcal{Q}_{-i}^\ell) - \delta_1). \tag{B.16}$$

To generalize (B.16) to all data of all subspaces, the following must hold for each $\ell = 1, ..., k$:

$$\mu(\mathcal{X}_\ell) + 3\rho + \delta_1 < (1 - 2\rho)\left(\min_{\{i : x_i \in X^{(\ell)}\}} r(\mathcal{Q}_{-i}^{(\ell)}) - \delta_1 \right). \tag{B.17}$$

This gives a first condition on δ and λ, which we call it "**dual separation condition**" under noise. Note that this reduces to exactly the geometric condition in [124]'s Theorem 2.5 when $\delta = 0$.

B.1.4 Avoid trivial solution

In this section we provide sufficient conditions on λ such that trivial solution $c = 0$, $e = x_i^{(\ell)}$ is not the optimal solution. For any optimal triplet (c, e, ν) we have $\nu = \lambda e$, a condition: $\|\nu\| < \lambda\|x_i^{(\ell)}\|$ implies that optimal $\|e\| < \|x_i^{(\ell)}\|$, so $e \neq x_i^{(\ell)}$. By the equality constraint, $X_{-i}^{(\ell)} c = x_i^{(\ell)} - e \neq 0$, therefore $\|c\|_1 \neq 0$. Now we will establish the condition on λ such that $\|\nu\| < \lambda\|x_i^{(\ell)}\|$.

An upper bound of $\|\nu\|$ and a lower bound of $\lambda\|x_i^{(\ell)}\|$ are readily available:

$$\begin{aligned}
\|\nu\| \leq \|\nu_1\| + \|\nu_2\| &\leq \frac{1}{r\big(\mathcal{Q}_{-i}^{\ell}\big) - \delta_1} + 2\lambda\delta\left(\frac{1}{r(\mathcal{Q}_{-i}^{\ell})} + 1\right)\left(1 + \frac{\delta}{r\big(\mathcal{Q}_{-i}^{\ell}\big) - \delta_1}\right) \\
&\leq \frac{1 + 3\lambda\delta + 2\lambda\delta^2}{r\big(\mathcal{Q}_{-i}^{\ell}\big) - \delta_1} + 2\lambda\delta, \\
\lambda\|x_i^{(\ell)}\| &\geq \lambda(\|y_i^{(\ell)}\| - \|z_i^{(\ell)}\|) \geq \lambda(1 - \delta).
\end{aligned}$$

So the sufficient condition on λ such that solution is non-trivial is

$$\frac{1 + 3\lambda\delta + 2\lambda\delta^2}{r\big(\mathcal{Q}_{-i}^{\ell}\big) - \delta_1} + 2\lambda\delta < \lambda(1 - \delta).$$

Reorganize the condition, we reach

$$\lambda > \frac{1}{(r\big(\mathcal{Q}_{-i}^{\ell}\big) - \delta_1)(1 - 3\delta) - 3\delta - 2\delta^2}. \tag{B.18}$$

For the inequality operations above to be valid, we need:

$$\begin{cases} r\big(\mathcal{Q}_{-i}^{\ell}\big) - \delta_1 > 0 \\ (r\big(\mathcal{Q}_{-i}^{\ell}\big) - \delta_1)(1 - 3\delta) - 3\delta - 2\delta^2 > 0 \end{cases}$$

Relax δ_1 to δ and solve the system of inequalities, we get:

$$\delta < \frac{3r + 4 - \sqrt{9r^2 + 20r + 16}}{2} = \frac{2r}{3r + 4 + \sqrt{9r^2 + 20r + 16}}.$$

Use $\sqrt{9r^2 + 20r + 16} \leq 3r + 4$ and impose the constraint for all $x_i^{(\ell)}$, we choose to

impose a stronger condition for every $\ell = 1, ..., L$:

$$\delta < \frac{\min_i r\big(\mathcal{Q}^{\ell}_{-i}\big)}{3\min_i r\big(\mathcal{Q}^{\ell}_{-i}\big) + 4}. \tag{B.19}$$

B.1.5 Existence of a proper λ

Basically, (B.17), (B.18) and (B.14) must be satisfied simultaneously for all $\ell = 1, ..., L$. Essentially (B.18) gives condition of λ from below, the other two each gives a condition from above. Denote $r_\ell := \min_{\{i:x_i \in X^{(\ell)}\}} r(\mathcal{Q}^{(\ell)}_{-i})$, $\mu_\ell := \mu(\mathcal{X}_\ell)$, the condition on λ is:

$$\begin{cases} \lambda & > \max_\ell \frac{1}{(r_\ell - \delta_1)(1-3\delta) - 3\delta - 2\delta^2} \\ \lambda & < \min_\ell \left(\frac{r_\ell - \mu_\ell - 2\delta_1}{\delta(1+\delta)(3+2r_\ell - 2\delta_1)} \vee \frac{2r_\ell}{\delta^2(r_\ell+1)} \right) \end{cases}$$

Note that on the left

$$\max_\ell \left\{ \frac{1}{(r_\ell - \delta_1)(1-3\delta) - 3\delta - 2\delta^2} \right\} = \frac{1}{(\max_\ell r_\ell - \delta_1)(1-3\delta) - 3\delta - 2\delta^2}.$$

On the right

$$\min_\ell \left\{ \frac{2r_\ell}{\delta^2(r_\ell+1)} \right\} = \frac{2\min_\ell r_\ell}{\delta^2(\min_\ell r_\ell + 1)}.$$

Denote $r = \min_\ell r_\ell$, it suffices to guarantee for each ℓ:

$$\begin{cases} \lambda & > \frac{1}{(r-\delta_1)(1-3\delta) - 3\delta - 2\delta^2} \\ \lambda & < \frac{r_\ell - \mu_\ell - 2\delta_1}{\delta(1+\delta)(3+2r_\ell - 2\delta_1)} \vee \frac{2r}{\delta^2(r+1)} \end{cases} \tag{B.20}$$

To understand this, when δ and μ is small then any λ values satisfying $\Theta(r) < \lambda < \Theta(r/\delta)$ will satisfy separation condition. We will now derive the condition on δ such that (B.20) is not an empty set.

B.1.6 Lower bound of break-down point

(B.19) gives one requirement on δ and the range of (B.20) being non-empty gives another. Combining these two leads to lower bound of the breakdown point. In other word, the algorithm will be robust to arbitrary corruptions with magnitude less than this point for some λ. Again, we relax δ_1 to δ in (B.20) to get:

$$\begin{cases} \frac{1}{(r-\delta)(1-3\delta)-3\delta-2\delta^2} < \frac{r_\ell-\mu_\ell-2\delta}{\delta(1+\delta)(3+2r_\ell-2\delta)} \\ \frac{1}{(r-\delta)(1-3\delta)-3\delta-2\delta^2} < \frac{2r}{\delta^2(r+1)}. \end{cases}$$

The first inequality in standard form is:

$$A\delta^3 + B\delta^2 + C\delta + D < 0 \quad \text{with}$$

$$\begin{cases} A = 0 \\ B = -(6r - r_\ell + 7 - \mu_\ell) \\ C = 3r_\ell r + 6r_\ell + 2r - 3\mu_\ell r + 3 - 4\mu_\ell \\ D = -r(r_\ell - \mu_\ell) \end{cases}$$

This is an extremely complicated 3^{rd} order polynomial. We will try to simplify it imposing a stronger condition. First extract and regroup μ_ℓ in first three terms, we get $(\delta^2 - 4\delta - 3r\delta)\mu_\ell$ which is negative, so we drop it. Second we express the remaining expression using:

$$f(r,\delta)\delta < r(r-\mu),$$

where $f(r,\delta) = -(6r - r_\ell + 7)\delta + 3r_\ell r + 6r_\ell + 2r + 2$. Note that since $\delta < 1$, we can write $f(r,\delta) \leq f(r,0) = 3r_\ell r + 6r_\ell + 2r + 2 \leq 3r_\ell^2 + 8r_\ell + 2$. Thus, a stronger condition on δ is established:

$$\delta < \frac{r(r_\ell - \mu_\ell)}{3r_\ell^2 + 8r_\ell + 2} \tag{B.21}$$

The second inequality in standard form is:

$$(1-r)\delta^2 + (6r^2+8r)\delta - 2r^2 < 0$$

By definition $r < 1$, we solve the inequality and get:

$$\begin{cases} \delta > \frac{-3r^2-4r-r\sqrt{9r^2+22r+18}}{1-r} \\ \delta < \frac{-3r^2-4r+r\sqrt{9r^2+22r+18}}{1-r} \end{cases}$$

The lower constraint is always satisfied. Rationalized the expression of the upper constraint, $1-r$ gets cancelled out:

$$\delta < \frac{2r^2}{3r^2+4r+r\sqrt{9r^2+22r+18}}.$$

It turns out that (B.19) is sufficient for the inequality to hold. This is by

$$\sqrt{9r^2+22r+18} < \sqrt{9r^2+24r+16} = 3r+4.$$

Combine with (B.21) we reach the overall condition:

$$\delta < \left\{ \frac{r(r_\ell - \mu_\ell)}{3r_\ell^2+8r_\ell+2} \right\} \vee \frac{r}{3r+4} = \frac{r(r_\ell-\mu_\ell)}{3r_\ell^2+8r_\ell+2}. \tag{B.22}$$

The first expression is always smaller because:

$$\frac{r}{3r+4} \geq \frac{rr_\ell}{3rr_\ell+4r_\ell} \geq \frac{rr_\ell}{3rr_\ell+4r_\ell+3r_\ell+2} \geq \frac{r(r_\ell-\mu_\ell)}{3r_\ell^2+8r_\ell+2}.$$

Verify that when (B.22) is true for all ℓ, there exists a single λ for solution of (3.2) to satisfy subspace detection property for all x_i. The proof of Theorem 3.1 is now complete.

B.2 Proof of Randomized Results

In this section, we provide proof to the Theorems about the three randomized models:

• **Determinitic data+random noise**

• **Semi-random data+random noise**

• **Fully random**

To do this, we need to bound δ_1, $\cos(\angle(z,\nu))$ and $\cos(\angle(y,\nu_2))$ when the Z follows *Random Noise Model*, such that a better dual separation condition can be obtained. Moreover, for *Semi-random* and *Random data model*, we need to bound $r(\mathcal{Q}_{-i}^{(\ell)})$ when data samples from each subspace are drawn uniformly and bound $\mu(\mathcal{X}_\ell)$ when subspaces are randomly generated.

These requires the following Lemmas.

Lemma B.4 (Upper bound on the area of spherical cap)**.** *Let $a \in \mathbb{R}^n$ be a random vector sampled from a unit sphere and z is a fixed vector. Then we have:*

$$Pr\left(|a^T z| > \epsilon\|z\|\right) \leq 2e^{\frac{-n\epsilon^2}{2}}$$

This Lemma is extracted from an equation in page 29 of Soltanolkotabi and Candes [124], which is in turn adapted from the upper bound on the area of spherical cap in Ball [6]. By definition of Random Noise Model, z_i has spherical symmetric, which implies that the direction of z_i distributes uniformly on an n-sphere. Hence Lemma B.4 applies whenever an inner product involves z.

As an example, , we write the following lemma

Lemma B.5 (Properties of Gaussian noise)**.** *For Gaussian random matrix $Z \in \mathbb{R}^{n\times N}$, if each entry $Z_{i,j} \sim N(0, \frac{\sigma}{\sqrt{n}})$, then each column z_i satisfies:*

1. $Pr(\|z_i\|^2 > (1+t)\sigma^2) \leq e^{\frac{n}{2}(\log(t+1)-t)}$
2. $Pr(|\langle z_i, z\rangle| > \epsilon\|z_i\|\|z\|) \leq 2e^{\frac{-n\epsilon^2}{2}}$

where z is any fixed vector(or random generated but independent to z_i).

Proof. The second property follows directly from Lemma B.4 as Gaussian vector has uniformly random direction.

To show the first property, we observe that the sum of n independent square Gaussian random variables follows χ^2 distribution with d.o.f n, in other word, we have

$$\|z_i\|^2 = |Z_{1i}|^2 + ... + |Z_{ni}|^2 \sim \frac{\sigma^2}{n}\chi^2(n).$$

By Hoeffding's inequality, we have an approximation of its CDF [44], which gives us

$$Pr(\|z_i\|^2 > \alpha\sigma^2) = 1 - \mathrm{CDF}_{\chi^2_n}(\alpha) \leq (\alpha e^{1-\alpha})^{\frac{n}{2}}.$$

Substitute $\alpha = 1 + t$, we get exactly the concentration statement. □

By Lemma B.5, $\delta = \max_i \|z_i\|$ is bounded with high probability. δ_1 has an even tighter bound because each $\mathcal{S}_\ell$ is low-rank. Likewise, $\cos(\angle(z, \nu))$ is bounded to a small value with high probability. Moreover, since $\nu = \lambda e = \lambda(x_i - X_{-i}c)$, $\nu_2 = \lambda\mathbb{P}_{\mathcal{S}_\ell^\perp}(z_i - Z_{-i}c)$, thus ν_2 is merely a weighted sum of random noise in a $(n - d_\ell)$-dimensional subspace. Consider y a fixed vector, $\cos(\angle(y, \nu_2))$ is also bounded with high probability.

Replace these observations into (B.7) and the corresponding bound of $\|\nu_1\|$ and $\|\nu_2\|$. We obtained the dual separation condition for under Random noise model.

Lemma B.6 (Dual separation condition under random noise)**.** *Let $\rho := \lambda\delta(1 + \delta)$ and*

$$\epsilon := \sqrt{\frac{6\log N + 2\log\max_\ell d_\ell}{n - \max_\ell d_\ell}} \leq \frac{C\log(N)}{\sqrt{n}}$$

for some constant C. Under random noise model, if for each $\ell = 1, ..., L$

$$\mu(\mathcal{X}_\ell) + 3\rho\epsilon + \delta\epsilon \leq (1 - 2\rho\epsilon)(\max_i r(\mathcal{Q}_{-i}^{(\ell)}) - \delta\epsilon),$$

then dual separation condition (B.7) *holds for all data points with probability at least* $1 - 7/N$.

Proof. Recall that we want to find an upper bound of $|\langle x, \nu \rangle|$.

$$|\langle x, \nu \rangle| \leq \mu \|\nu_1\| + \|y\| \|\nu_2\| |\cos(\angle(y, \nu_2))| + \|z\| \|\nu\| |\cos(\angle(z, \nu))| \tag{B.23}$$

Here we will bound the two cosine terms and δ_1 under random noise model.

As discussed above, directions of z and ν_2 are independently and uniformly distributed on the n-sphere. Then by Lemma B.4,

$$\begin{cases} Pr\left(\cos(\angle(z, \nu)) > \sqrt{\frac{6 \log N}{n}}\right) \leq \frac{2}{N^3} \\ Pr\left(\cos(\angle(y, \nu_2)) > \sqrt{\frac{6 \log N}{n - d_\ell}}\right) \leq \frac{2}{N^3} \\ Pr\left(\cos(\angle(z, \nu_2)) > \sqrt{\frac{6 \log N}{n}}\right) \leq \frac{2}{N^3} \end{cases}$$

Using the same technique, we provide a bound for δ_1. Given orthonormal basis U of S_ℓ, $\mathbb{P}_{S_\ell} z = U U^T z$, then

$$\|U U^T z\| = \|U^T z\| \leq \sum_{i=1,\ldots,d_\ell} |U_{:,i}^T z|.$$

Apply Lemma B.4 for each i , then apply union bound, we get:

$$Pr\left(\|\mathbb{P}_{S_\ell} z\| > \sqrt{\frac{2 \log d_\ell + 6 \log N}{n}} \delta\right) \leq \frac{2}{N^3}$$

Since δ_1 is the worse case bound for all L subspace and all N noise vector, then a union bound gives:

$$Pr\left(\delta_1 > \sqrt{\frac{2 \log d_\ell + 6 \log N}{n}} \delta\right) \leq \frac{2L}{N^2}$$

Moreover, we can find a probabilistic bound for $\|\nu_1\|$ too by a random variation of (B.9)

which is now

$$y_i^T \nu_1 + (\mathbb{P}_{\mathcal{S}_\ell} z_i)^T \nu_1 \leq 1 - z_i^T \nu_2 \leq 1 + \delta_2 \|\nu_2\| |\cos \angle(z_i, \nu_2)|. \tag{B.24}$$

Substituting the upper bound of the cosines, we get:

$$|\langle x, \nu \rangle| \leq \mu \|\nu_1\| + \|y\| \|\nu_2\| \sqrt{\frac{6 \log N}{n - d_\ell}} + \|z\| \|\nu\| \sqrt{\frac{6 \log N}{n}}$$

$$\|\nu_1\| \leq \frac{1 + \delta \|\nu_2\| \sqrt{\frac{6 \log N}{n}}}{r(\mathcal{Q}_{-i}^\ell) - \delta_1}, \qquad \|\nu_2\| \leq 2\lambda\delta \left(\frac{1}{r(\mathcal{Q}_{-i}^\ell)} + 1 \right)$$

Denote $r := r(\mathcal{Q}_{-i}^\ell)$, $\epsilon := \sqrt{\frac{6 \log N + 2 \log \max_\ell d_\ell}{n - \max_\ell d_\ell}}$ and $\mu := \mu(\mathcal{X}_\ell)$ we can further relax the bound into

$$\begin{aligned} |\langle x, \nu \rangle| \leq & \frac{\mu + \delta\epsilon}{r - \epsilon\delta} + \frac{(\mu + \delta\epsilon) 2\delta^2 \epsilon}{r - \epsilon\delta} \left(\frac{1}{r} + 1 \right) + 2\lambda\delta\epsilon \left(\frac{1}{r} + 1 \right) + 2\lambda\delta^2\epsilon \left(\frac{1}{r} + 1 \right) \\ \leq & \frac{\mu + \delta\epsilon + 3\lambda\delta(1+\delta)\epsilon}{r - \epsilon\delta} + 2\lambda\delta(1+\delta)\epsilon. \end{aligned}$$

Note that here in order to get rid of the higher order term $\frac{1}{r(r-\epsilon\delta)}$, we used $\delta < r$ and $\mu + \delta\epsilon < 1$ to construct $\frac{(\mu+\delta\epsilon)\delta^2\epsilon}{r(r-\delta\epsilon)} < \frac{\delta\epsilon}{r-\delta\epsilon}$ as in the proof of Theorem 3.1. Now impose the dual detection constraint on the upper bound, we get:

$$2\lambda\delta(1+\delta)\epsilon + \frac{\mu + \delta\epsilon + 3\lambda\delta(1+\delta)\epsilon}{r - \delta\epsilon} < 1.$$

Replace $\rho := \lambda\delta(1+\delta)$ and reorganize the inequality, we reach the desired condition:

$$\mu + 3\rho\epsilon + \delta\epsilon \leq (1 - 2\rho\epsilon)(r - \delta\epsilon).$$

There are N^2 instances for each of the three events related to the consine value, apply union bound we get the failure probability $\frac{6}{N} + \frac{2L}{N^2} \leq \frac{7}{N}$. This concludes the proof. $\square$

B.2.1 Proof of Theorem 3.2

Lemma B.6 has already provided the separation condition. The things left are to find the range of λ and update the condition of δ.

The range of λ: Follow the same arguments in Section B.1.4 and Section B.1.5, re-derive the upper bound from the relationship in Lemma B.6 and substitute the tighter bound of δ_1 where applicable. Again let $r_\ell = \min_i r(Q^\ell_{-i})$, $\mu_\ell = \mu(\mathcal{X}_\ell)$ and $r = \min_\ell r_\ell$. We get the range of λ under random noise model:

$$\begin{cases} \lambda > \dfrac{1}{(r-\delta\epsilon)(1-3\delta)-3\delta-2\delta^2} \\ \lambda < \min\limits_{\ell=1,\dots,L} \left\{ \dfrac{r_\ell - \mu_\ell - 2\delta\epsilon}{\epsilon\delta(1+\delta)(3+2r_\ell-2\delta\epsilon)} \right\} \vee \dfrac{2r}{\delta^2(r+1)} \end{cases} \tag{B.25}$$

Remark B.1. *A critical difference from the deterministic noise model is that now under the paradigm of small μ and δ, if $\delta > \epsilon$, the second term in the upper bound is actually tight. Then the valid range of λ is expanded an order to $\Theta(1/r) \leq \lambda < \Theta(r/\delta^2)$.*

The condition of δ: Re-derive (B.19) using $\delta_1 \leq \epsilon\delta$, we get:

$$\delta < \frac{r}{3r+3+\epsilon} \tag{B.26}$$

Likewise, we re-derive (B.21) from the new range of λ in (B.25). The first inequality in standard form is,

$$A\delta^3 + B\delta^2 + C\delta + D < 0 \quad \text{with}$$

$$\begin{cases} A = 6\epsilon^2 - 6\epsilon, \\ B = -(3\epsilon + 4\epsilon^2 + \epsilon r_\ell - 2r_\ell + 6\epsilon r + 2\mu_\ell - 3\mu_\ell\epsilon), \\ C = 3r_\ell r + 3r_\ell + 3\epsilon r_\ell + 3\epsilon + 2\epsilon r - 3\mu_\ell r - 3\mu_\ell - \epsilon\mu_\ell, \\ D = -r(r_\ell - \mu_\ell), \end{cases}$$

apply the same trick of removing the negative μ term and define

$$f(r,\delta) := A\delta^2 + B\delta + C$$

such that the 3^{rd}-order polynomial inequality becomes $f(r,\delta)\delta < r(r_\ell - \mu_\ell)$.Rearrange the expressions and drop negative terms, we get

$$\begin{aligned}
& f(r,\delta) < B\delta + C \\
=& -\left[3\epsilon + 4\epsilon^2 + 2\epsilon(r_\ell - \mu_\ell) + 6\epsilon r\right]\delta + 2(r_\ell - \mu_\ell)\delta \\
& + [3(r_\ell - \mu_\ell)r + 3(r_\ell - \mu_\ell) + 3\epsilon(r_\ell - \mu_\ell) + 2\epsilon r + 3\epsilon] \\
& + (r_\ell - \mu_\ell)\epsilon\delta + 2\mu_\ell\epsilon\delta - \mu_\ell\epsilon \\
<& 3(r_\ell - \mu_\ell)r + 5(r_\ell - \mu_\ell) + 4\epsilon(r_\ell - \mu_\ell) + 2\epsilon r + 3\epsilon.
\end{aligned}$$

Therefore, a sufficient condition of δ is

$$\delta < \frac{r(r_\ell - \mu_\ell)}{3(r_\ell - \mu_\ell)r + 5(r_\ell - \mu_\ell) + 4\epsilon(r_\ell - \mu_\ell) + 2\epsilon r + 3\epsilon}. \tag{B.27}$$

When $r > r_\ell - \mu_\ell$, we have $(r_\ell - \mu_\ell)/r < 1$. Then

$$\text{(B.27)} \Leftarrow \quad \delta < \frac{r_\ell - \mu_\ell}{3(r_\ell - \mu_\ell) + 5 + \epsilon(4 + 2 + 3/r)} \Leftarrow \quad \delta < \frac{r_\ell - \mu_\ell}{3r + 5 + \epsilon(6 + 3/r)}.$$

When $r < r_\ell - \mu_\ell$, we have $r/(r_\ell - \mu_\ell) < 1$. Since $r < r_\ell$,

$$\text{(B.27)} \Leftarrow \quad \delta < \frac{r}{3r + 5 + \epsilon(4 + 2 + 3/(r_\ell - \mu_\ell))} \Leftarrow \quad \delta < \frac{r}{3r + 5 + \epsilon(6 + 3/r)}$$

Combining the two cases, we have:

$$\delta < \frac{\min\{r, r_\ell - \mu_\ell\}}{3r + 5 + \epsilon(6 + 3/r)} \tag{B.28}$$

For the second inequality, the quadratic polynomial is now

$$(1 + 5r - 6r\epsilon)\delta^2 + (6r^2 + 2\epsilon r + 6r)\delta - 2r^2 < 0.$$

Check that $1 + 5r - 6r\epsilon > 0$. We solve the quadratic inequality and get a slightly

stronger condition than (B.26), which is

$$\delta < \frac{r}{3r + 4 + \epsilon}. \tag{B.29}$$

Note that (B.28) $\Rightarrow$ (B.29), so (B.28) alone is sufficient. In fact, when $\epsilon(6r + 3)/r < 1$ or equivalently $r > 3\epsilon/(1 - 6\epsilon)$, which are almost always true, a neater expression is:

$$\delta < \frac{\min\{r, r_\ell - \mu_\ell\}}{3r + 6}.$$

Finally, as the condition needs to be satisfied for all ℓ, the output of the $\min$ function at the smallest bound is always $r_\ell - \mu_\ell$. This observation allows us to replace $\min\{r, r_\ell - \mu_\ell\}$ with simple $(r_\ell - \mu_\ell)$, which concludes the proof for Theorem 3.2.

B.2.2 Proof of Theorem 3.3

To prove Theorem 3.3, we only need to bound inradii r and incoherence parameter μ under the new assumptions, then plug into Theorem 3.2.

Lemma B.7 (Inradius bound of random samples)**.** *In random sampling setting, when each subspace is sampled $N_\ell = \kappa_\ell d_\ell$ data points randomly, we have:*

$$Pr\left\{ c(\kappa_\ell)\sqrt{\frac{\beta \log(\kappa_\ell)}{d_\ell}} \leq r(\Omega_{-i}^{(\ell)}) \textit{ for all pairs } (\ell, i) \right\} \geq 1 - \sum_{\ell=1}^{L} N_\ell e^{-d_\ell^\beta N_\ell^{1-\beta}}$$

This is extracted from Section-7.2.1 of Soltanolkotabi and Candes [124]. $\kappa_\ell = (N_\ell - 1)/d_\ell$ is the relative number of iid samples. $c(\kappa)$ is some positive value for all $\kappa > 1$ and for a numerical value κ_0, if $\kappa > \kappa_0$, we can take $c(\kappa) = \frac{1}{\sqrt{8}}$. Take $\beta = 0.5$, we get the required bound of r in Theorem 3.3.

Lemma B.8 (Incoherence bound)**.** *In deterministic subspaces/random sampling setting,*

the subspace incoherence is bounded from above:

$$Pr\Big\{\mu(\mathcal{X}_\ell) \le t\left(\log[(N_{\ell_1}+1)N_{\ell_2}] + \log L\right)\frac{\text{aff}(S_{\ell_1}, S_{\ell_2})}{\sqrt{d_{\ell_1}}\sqrt{d_{\ell_2}}} \\ \textit{for all pairs}(\ell_1,\ell_2)\ \textit{with}\ \ell_1 \neq \ell_2\Big\} \ge 1 - \frac{1}{L^2}\sum_{\ell_1\neq\ell_2}\frac{1}{(N_{\ell_1}+1)N_{\ell_2}}e^{-\frac{t}{4}}$$

Proof of Lemma B.8. The proof is an extension of the same proof in Soltanolkotabi and Candes [124]. First we will show that when noise $z_i^{(\ell)}$ is spherical symmetric, and clean data points $y_i^{(\ell)}$ has iid uniform random direction, projected dual directions $v_i^{(\ell)}$ also follows uniform random distribution.

Now we will prove the claim. First by definition,

$$v_i^{(\ell)} = v(x_i^{(\ell)}, X_{-i}^{(\ell)}, \mathcal{S}_\ell, \lambda) = \frac{\mathbb{P}_{S_\ell}\nu}{\|\mathbb{P}_{S_\ell}\nu\|} = \frac{\nu_1}{\|\nu_1\|}.$$

ν is the unique optimal solution of $\mathbf{D}_1$ (B.5). Fix λ, $\mathbf{D}_1$ depends on two inputs, so we denote $\nu(x, X)$ and consider ν a function. Moreover, $\nu_1 = \mathbb{P}_{\mathcal{S}}\nu$ and $\nu_2 = \mathbb{P}_{\mathcal{S}^\perp}\nu$. Let $U \in n \times d$ be a set of orthonormal basis of d-dimensional subspace $\mathcal{S}$ and a rotation matrix $R \in \mathbb{R}^{d\times d}$. Then rotation matrix within subspace is hence URU^T.

$$x_1 := \mathbb{P}_{\mathcal{S}}x = y + z_1 \sim URU^T y + URU^T z_1$$
$$x_2 := \mathbb{P}_{\mathcal{S}^\perp}x = z_2$$

As y is distributed uniformly on unit sphere of $\mathcal{S}$, and z is spherical symmetric noise(hence z_1 and z_2 are also spherical symmetric in subspace), for any fixed $\|x_1\|$, the distribution is uniform on the sphere. It suffices to show the uniform distribution of ν_1 with fixed $\|x_1\|$.

Since inner product $\langle x, \nu\rangle = \langle x_1, \nu_1\rangle + \langle x_2, \nu_2\rangle$, we argue that if ν is optimal solution of

$$\max_\nu \langle x, \nu\rangle - \frac{1}{2\lambda}\nu^T\nu, \quad \text{subject to:} \quad \|X^T\nu\|_\infty \le 1,$$

then the optimal solution of R-transformed optimization

$$\begin{aligned}&\max_{\nu} \langle URU^T x_1 + x_2, \nu\rangle - \frac{1}{2\lambda}\nu^T\nu,\\ &\text{subject to:} \quad \|(URU^T X_1 + X_2)^T \nu\|_\infty \leq 1,\end{aligned}$$

is merely the transformed ν under the same R:

$$\begin{aligned}\nu(R) &= \nu(URU^T x_1 + x_2, URU^T X_1 + X_2)\\ &= URU^T \nu_1(x, X) + \nu_2(x, X) = URU^T \nu_1 + \nu_2.\end{aligned} \tag{B.30}$$

To verify the argument, check that $\nu^T\nu = \nu(R)^T\nu(R)$ and

$$\langle URU^T x_1 + x_2, \nu(R)\rangle = \langle URU^T x_1, URU^T \nu_1\rangle + \langle x_1, \nu_2\rangle = \langle x, \nu\rangle$$

for all inner products in both objective function and constraints, preserving the optimality.

By projecting (B.30) to subspace, we show that operator $v(x, X, S)$ is linear *vis a vis* subspace rotation URU^T, i.e.,

$$v(R) = \frac{\mathbb{P}_{S_\ell}\nu(R)}{\|\mathbb{P}_{S_\ell}\nu(R)\|} = \frac{URU^T\nu_1}{\|URU^T\nu_1\|} = URU^T v. \tag{B.31}$$

On the other hand, we know that

$$v(R) = v(URU^T x_1 + x_2, URU^T X_1 + X_2, \mathcal{S}) \sim v(x, X, \mathcal{S}), \tag{B.32}$$

where $A \sim B$ means that the random variables A and B follows the same distribution. When $\|x_1\|$ is fixed and each columns in X_1 has fixed magnitudes, $URU^T x_1 \sim x_1$ and $URU^T X_1 \sim X_1$. Since (x_1, X_1) and (x_2, X_2) are independent, we can also marginalize out the distribution of x_2 and X_2 by considering fixed (x_2, X_2). Combining (B.31)

and (B.32), we conclude that for any rotation R,

$$v_i^{(\ell)}(R) \sim URU^T v_i^{(\ell)}.$$

Now integrate the marginal probability of $v_i^{(\ell)}$ over $\|x_{i1}^{\ell}\|$, every column's magnitude of $X_{-i_1}^{\ell}$ and all (x_2, X_2), we showed that the overall distribution of $v_i^{(\ell)}$ is indeed uniformly distributed in the unit sphere of $\mathcal{S}$.

After this key step, the rest is identical to Lemma 7.5 of Soltanolkotabi and Candes [124]. The idea is to use Lemma B.4(upper bound of area of spherical caps) to bound pairwise inner product and Borell's inequality to bound the deviation from expected consine canonical angles, namely, $\|{U^{(k)}}^T U^{(\ell)}\|_F/\sqrt{d_\ell}$. □

B.2.3 Proof of Theorem 3.4

The proof of this theorem is also an invocation of Theorem 3.2 with specific inradii bound and incoherence bound. The bound of inradii is exactly Lemma B.7 with $\beta = 0.5$, $\kappa_\ell = \kappa$, $d_\ell = d$. The bound of incoherence is given by the following Lemma that is extracted from Step 2 of Section 7.3 in Soltanolkotabi and Candes [124].

Lemma B.9 (Incoherence bound of random subspaces)**.** *In random subspaces setting, the projected subspace incoherence is bounded from above:*

$$Pr\left\{\mu(\mathcal{X}_\ell) \leq \sqrt{\frac{6\log N}{n}} \textit{ for all } \ell\right\} \geq 1 - \frac{2}{N}.$$

Now that we have shown that projected dual directions are randomly distributed in their respective subspace, as the subspaces themselves are randomly generated, all clean data points y and projected dual direction v from different subspaces can be considered iid generated from the ambient space. The proof of Lemma B.9 follows by simply applying Lemma B.4 and union bound across all N^2 events.

By plug in these expressions into Theorem 3.2, we showed that it holds with high probability as long as the conditions in Theorem 3.4 is true.

B.3 Geometric interpretations

In this section, we attempt to give some geometric interpretation of the problem so that the results stated in this chapter can be better understood and at the same time, reveal the novelties of our analysis over Soltanolkotabi and Candes [124]. All figures in this section are drawn with "geom3d" [92] and "GBT7.3" [138] in Matlab.

We start with an illustration of the projected dual direction in contrast to the original dual direction[124].

Dual direction v.s. Projected dual direction:

An illustration of original dual direction is given in Figure B.1 for data point y. The

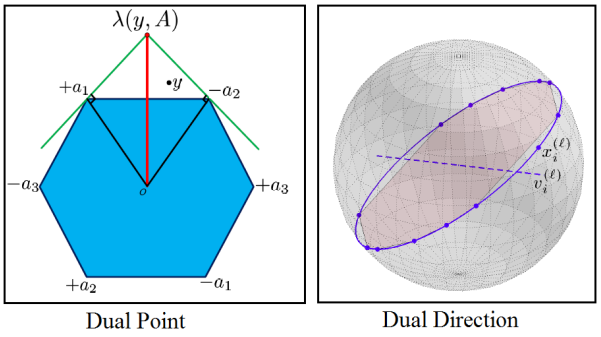


Figure B.1: The illustration of dual direction in Soltanolkotabi and Candes [124].

projected dual direction can be easier understood algebraically. By definition, it is the projected optimal solution of (B.5) to the true subspace. To see it more clearly, we plot the feasible region of ν in Figure B.2 (b), and the projection of the feasible region in Figure B.3. As (B.5) is not an LP (it has a quadratic term in the objective function), projected dual direction cannot be easily determined geometrically as in Figure B.1. Nevertheless, it turns out to be sufficient to know the feasible region and the optimality of the solution.

Magnitude of dual variable ν:

A critical step of our proof is to bound the magnitude of $\|\nu_1\|$ and $\|\nu_2\|$. This is a simple task in the noiseless case as Soltanolkotabi and Candes merely take the circumradius of the full feasible region as a bound. This is sufficient because the feasible

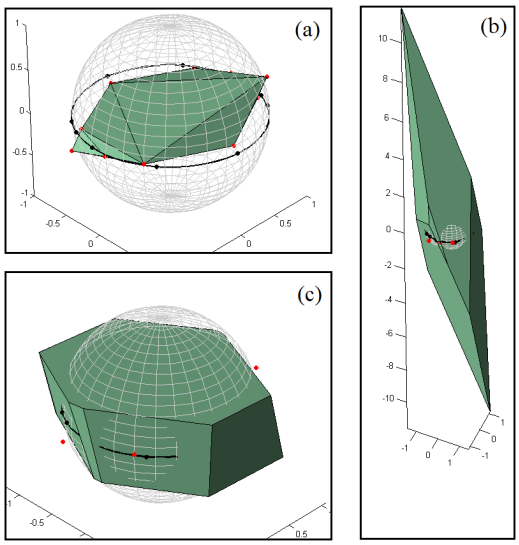


Figure B.2: Illustration of **(a)** the convex hull of noisy data points, **(b)** its polar set and **(c)** the intersection of polar set and $\|\nu_2\|$ bound. The polar set (b) defines the feasible region of (B.5). It is clear that ν_2 can take very large value in (b) if we only consider feasibility. By considering optimality, we know the optimal ν must be inside the region in (c).

region is a cylinder perpendicular to the subspace and there is no harm choosing only solutions within the intersection of the cylinder and the subspace. Indeed, in noiseless case, we can choose arbitrary ν_2 because $Y^T(\nu_1 + \nu_2) = Y^T\nu_1$.

In the noisy case however, the problem becomes harder. Instead of a cylinder, the feasible region is now a spindle shaped polytope (see Figure B.2(b)) and the choice of ν_2 has an impact on the objective value. That is why we need to consider the optimality condition and give $\|\nu_2\|$ a bound.

In fact, noise may tilt the direction of the feasible region (especially when the noise is adversarial). As $\|\nu_2\|$ grows, $\|\nu_1\|$ can potentially get large too. Our bound of $\|\nu_1\|$ reflects precisely the case as it is linearly dependent on $\|\nu_2\|$ (see (B.11)). We remark that in the case of random noise, the dependency on $\|\nu_2\|$ becomes much weaker (see the proof of Lemma B.6).

Geometrically, the bound of ν_2 can be considered a cylinder[1] (ℓ_2 constrained in the $\mathcal{S}^\perp$ and unbounded in $\mathcal{S}$ subspace) that intersect the spindle shaped feasible region, so

[1] In the simple illustration, the cylinder is in fact just the sandwich region $|z| \leq$ some bound.

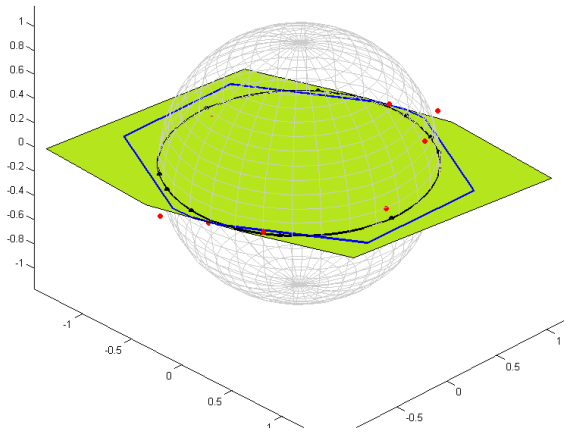


Figure B.3: The projection of the polar set (the green area) in comparison to the projection of the polar set with $\|\nu_2\|$ bound (the blue polygon). It is clear that the latter is much smaller.

that we know the optimal ν may never be at the tips of the spindle (see Figure B.2 and B.3). Algebraically, we can consider this as an effect of the quadratic penalty term of ν in the (B.5).

The guarantee in Theorem 3.1:

The geometric interpretation and comparison of the noiseless guarantee and our noisy guarantee are given earlier in Figure 3.4. Geometrically, noise reduces the successful region (the solid blue polygon) in two ways. One is subtractive, in a sense that the inradius is smaller (see the bound of $\|\nu_1\|$); the other is multiplicative, as the entire successful region shrinks with a factor related to noise level (something like $1 - f(\delta)$). Readers may refer to (B.16) for an algebraic point of view.

The subtractive effect can also be interpreted in the robust optimization point of view, where the projection of every points inside the uncertainty set (the red balls in Figure 3.4) must fall into the successful region (the dashed red polygon). Either way, it is clear that the error Lasso-SSC can provably tolerate is proportional to the geometric gap $r - \mu$ given in the noiseless case.

B.4 Numerical algorithm to solve Matrix-Lasso-SSC

In this section we outline the steps of solving the matrix version of Lasso-SSC below ((3.3) in the Chapter 3)

$$\begin{aligned} \min_{C} \ & \|C\|_1 + \frac{\lambda}{2}\|X - XC\|_F^2 \\ \text{s.t.} \quad & \operatorname{diag}(C) = 0, \end{aligned} \tag{B.33}$$

While this convex optimization can be solved by some off-the-shelf general purpose solver such as CVX, such approach is usually slow and non-scalable. An ADMM [17] version of the problem is described here for fast computation. It solves an equivalent optimization program

$$\begin{aligned} \min_{C} \ & \|C\|_1 + \frac{\lambda}{2}\|X - XJ\|_F^2 \\ \text{s.t.} \quad & J = C - \operatorname{diag}(C). \end{aligned} \tag{B.34}$$

We add to the Lagrangian with an additional quadratic penalty term for the equality constraint and get the augmented Lagrangian

$$\begin{aligned} \mathcal{L} =& \|C\|_1 + \frac{\lambda}{2}\|X - XJ\|_F^2 + \frac{\mu}{2}\|J - C + \operatorname{diag}(C)\|_F^2 \\ & + tr(\Lambda^T(J - C + \operatorname{diag}(C))), \end{aligned}$$

where Λ is the dual variable and μ is a parameter. Optimization is done by alternatingly optimizing over J, C and Λ until convergence. The update steps are derived by solving $\partial\mathcal{L}/\partial J = 0$ and $\partial\mathcal{L}/\partial C = 0$, it's non-differentiable for C at origin so we use the now standard soft-thresholding operator[47]. For both variables, the solution is in closed-form. For the update of Λ, it is simply gradient descent. For details of the ADMM algorithm and its guarantee, please refer to Boyd et al. [17]. To accelerate the convergence, it is possible to introduce a parameter ρ and increase μ by $\mu = \rho\mu$ at every iteration. The full algorithm is summarized in Algorithm 6.

Note that for the special case when $\rho = 1$, the inverse of $(\lambda Y^T Y + \mu I)$ can be pre-computed, such that the iteration is linear time. Empirically, we found it good to set $\mu = \lambda$ and it takes roughly 50-100 iterations to converge to a sufficiently good

Algorithm 6 Matrix-Lasso-SSC

Input: Data points as columns in $X \in \mathbb{R}^{n \times N}$, tradeoff parameter λ, numerical parameters μ_0 and ρ.
Initialize $C = 0$, $J = 0$, $\Lambda = 0$, $k = 0$.
while not converged **do**

1. Update J by $J = (\lambda X^T X + \mu_k I)^{-1}(\lambda X^T X + \mu_k C - \Lambda)$.
2. Update C by $C^{'} = \text{SoftThresh}_{\frac{1}{\mu_k}} \left(J + \Lambda/\mu_k\right), C = C^{'} - \text{diag}(C^{'})$.
3. Update Λ by $\Lambda = \Lambda + \mu_k(J - C)$
4. Update parameter $\mu_{k+1} = \rho\mu_k$.
5. Iterate $k = k + 1$;

end while
Output: Affinity matrix $W = |C| + |C|^T$

points. We remark that the matrix version of the algorithm is much faster than column-by-column ADMM-Lasso especially for the cases when $N > n$. See the experiments.

We would like to point out that Elhamifar and Vidal [57] had formulated a more general version of SSC to account for not only noisy but also sparse corruptions in the Appendix of their arxiv paper while we were preparing for submission. The ADMM algorithm for Matrix-Lasso-SSC described here can be considered as a special case of the Algorithm 2 in their paper.

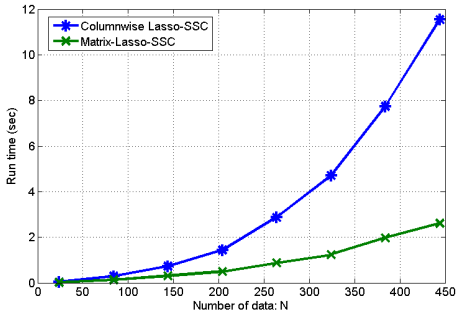


Figure B.4: Run time comparison with increasing number of data. Simulated with $n = 100, d = 4, L = 3, \sigma = 0.2$, κ increases from 2 to 40 such that the number of data goes from 24- 480. It appears that the matrix version scales better with increasing number of data compared to columnwise LASSO.

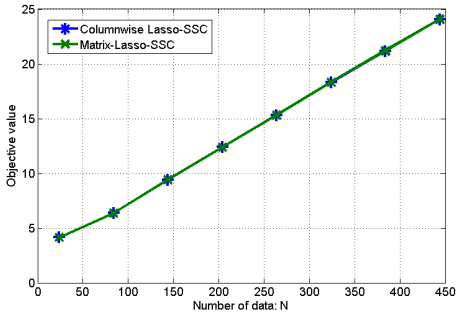


Figure B.5: Objective value comparison with increasing number of data. Simulated with $n = 100, d = 4, L = 3, \sigma = 0.2$, κ increases from 2 to 40 such that the number of data goes from 24-480. The objective value obtained at stop points of two algorithms are nearly the same.

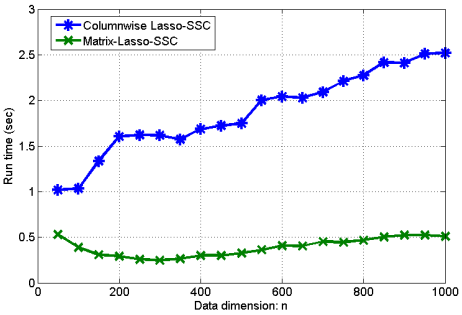


Figure B.6: Run time comparison with increasing dimension of data. Simulated with $\kappa = 5, d = 4, L = 3, \sigma = 0.2$, ambient dimension n increases from 50 to 1000. Note that the dependence on dimension is weak at the scale due to the fast vectorized computation. Nevertheless, it is clear that the matrix version of SSC runs faster.

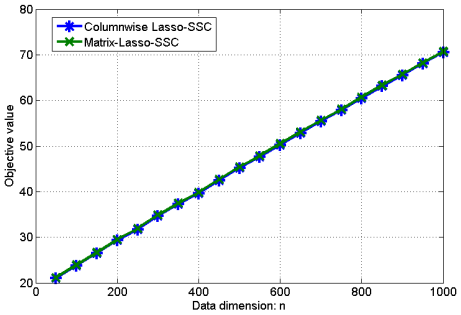


Figure B.7: Objective value comparison with increasing dimension of data. Simulated with $\kappa = 5, d = 4, L = 3, \sigma = 0.2$, ambient dimension n increases from 50 to 1000. The objective value obtained at stop points of two algorithms are nearly the same.

Appendix C

Appendices for Chapter 4

C.1 Proof of Theorem 4.1 (the deterministic result)

Theorem 4.1 is proven by duality. As described in the main text, it involves constructing two levels of fictitious optimizations. For convenience, we illustrate the proof with only three subspaces. Namely, $X = [X^{(1)}X^{(2)}X^{(3)}]$ and $\mathcal{S}_1$ $\mathcal{S}_2$ $\mathcal{S}_3$ are all d-dimensional subspaces. Having more than 3 subspaces and subspaces of different dimensions are perfectly fine and the proof will be the same.

C.1.1 Optimality condition

We start by describing the subspace projection critical in the proof of matrix completion and RPCA[25, 27]. We need it to characterize the subgradient of nuclear norm.

Define projection $\mathcal{P}_T$ (and $\mathcal{P}_{T^\perp}$) to both column and row space of low-rank matrix C (and its complement) as

$$\mathcal{P}_T(X) = UU^TX + XVV^T - UU^TXVV^T,$$

$$\mathcal{P}_{T^\perp}(X) = (I - UU^T)X(I - VV^T),$$

where UU^T and VV^T are projections matrix defined from skinny SVD of $C = U\Sigma V^T$.

Lemma C.1 (Properties of $\mathcal{P}_T$ and $\mathcal{P}_{T^\perp}$)**.**

$$\langle \mathcal{P}_T(X), Y\rangle = \langle X, \mathcal{P}_T(Y)\rangle = \langle \mathcal{P}_T(X), \mathcal{P}_T(Y)\rangle$$

$$\langle \mathcal{P}_{T^\perp}(X), Y\rangle = \langle X, \mathcal{P}_{T^\perp}(Y)\rangle = \langle \mathcal{P}_{T^\perp}(X), \mathcal{P}_{T^\perp}(Y)\rangle$$

Proof. Using the property of inner product $\langle X, Y\rangle = \langle X^T, Y^T\rangle$ and definition of adjoint operator $\langle AX, Y\rangle = \langle X, A^*Y\rangle$, we have

$$\begin{aligned}
\langle \mathcal{P}_T(X), Y\rangle &= \langle UU^TX, Y\rangle + \langle XVV^T, Y\rangle - \langle UU^TXVV^T, Y\rangle \\
&= \langle UU^TX, Y\rangle + \langle VV^TX^T, Y^T\rangle - \langle VV^TX^T, (UU^TY)^T\rangle \\
&= \langle X, UU^TY\rangle + \langle X^T, VV^TY^T\rangle - \langle X^T, VV^TY^TUU^T\rangle \\
&= \langle X, UU^TY\rangle + \langle X, YVV^T\rangle - \langle X, UU^TYVV^T\rangle = \langle X, \mathcal{P}_T(Y)\rangle.
\end{aligned}$$

Use the equality with $X = X, Y = \mathcal{P}_T(Y)$, we get

$$\langle X, \mathcal{P}_T(\mathcal{P}_T(Y))\rangle = \langle \mathcal{P}_T(X), \mathcal{P}_T(Y)\rangle.$$

The result for $\mathcal{P}_{T^\perp}$ is the same as the third term in the previous derivation as $I - UU^T$ and $I - VV^T$ are both projection matrices that are self-adjoint. □

In addition, given index set D, we define projection $\mathcal{P}_D$, such that

$$\mathcal{P}_D(X) = \begin{cases} [\mathcal{P}_D(X)]_{ij} = X_{ij}, & \text{if } (i,j) \in D; \\ [\mathcal{P}_D(X)]_{ij} = 0, & \text{Otherwise.} \end{cases}$$

For example, when $D = \{(i,j)|i = j\}$, $\mathcal{P}_D(X) = 0 \Leftrightarrow \text{diag}(X) = 0$.

Consider general convex optimization problem

$$\min_{C_1, C_2} \|C_1\|_* + \lambda\|C_2\|_1 \quad s.t. \quad B = AC_1, \quad C_1 = C_2, \quad \mathcal{P}_D(C_1) = 0 \qquad \text{(C.1)}$$

where $A \in R^{n\times m}$ is arbitrary dictionary and $B \in R^{n\times N}$ is data samples. Note that

when $B = X$, $A = X$, (C.1) is exactly (4.1).

Lemma C.2. *For optimization problem* (C.1)*, if we have a quadruplet* $(C, \Lambda_1, \Lambda_2, \Lambda_3)$ *where* $C_1 = C_2 = C$ *is feasible,* $\mathrm{supp}(C) = \Omega \subseteq \tilde{\Omega}$, $\mathrm{rank}(C) = r$ *and skinny SVD of* $C = U\Sigma V^T$ *(*Σ *is an* $r \times r$ *diagonal matrix and* U, V *are of compatible size), moreover if* Λ_1, Λ_2, Λ_3 *satisfy*

$$\begin{array}{lll} \text{① } \mathcal{P}_T(A^T\Lambda_1 - \Lambda_2 - \Lambda_3) = UV^T & \text{③ } [\Lambda_2]_\Omega = \lambda \mathrm{sgn}([C]_\Omega) & \text{⑤ } [\Lambda_2]_{\tilde{\Omega}^c} < \lambda \\ \text{② } \|\mathcal{P}_{T^\perp}(A^T\Lambda_1 - \Lambda_2 - \Lambda_3)\| \leq 1 & \text{④ } [\Lambda_2]_{\Omega^c \bigcap \tilde{\Omega}} \leq \lambda & \text{⑥ } \mathcal{P}_{D^c}(\Lambda_3) = 0 \end{array}$$

then all optimal solutions to (C.1) *satisfy* $\mathrm{supp}(C) \subseteq \tilde{\Omega}$.

Proof. The subgradient of $\|C\|_*$ is $UV^T + W_1$ for any $W_1 \in T^\perp$ and $\|W_1\| \leq 1$. For any optimal solution C^* we may choose W_1 such that $\|W_1\| = 1$, $\langle W_1, \mathcal{P}_{T^\perp} C^* \rangle = \|\mathcal{P}_{T^\perp} C^*\|_*$. Then by the definition of subgradient, convex function $\|C\|_*$ obey

$$\begin{aligned} \|C^*\|_* &\geq \|C\|_* + \langle UV^T + W_1, C^* - C \rangle \\ &= \langle UV^T, \mathcal{P}_T(C^* - C) \rangle + \langle UV^T, \mathcal{P}_{T^\perp}(C^* - C) \rangle + \langle W_1, C^* - C \rangle \\ &= \langle UV^T, \mathcal{P}_T(C^* - C) \rangle + \|\mathcal{P}_{T^\perp} C^*\|_*. \end{aligned} \tag{C.2}$$

To see the equality, note that $\langle UV^T, \mathcal{P}_{T^\perp}(A) \rangle = 0$ for any compatible matrix A and the following identity that follows directly from the construction of W_1 and Lemma C.1

$$\langle W_1, C^* - C \rangle = \langle \mathcal{P}_{T^\perp} W_1, C^* - C \rangle = \langle W_1, \mathcal{P}_{T^\perp}(C^* - C) \rangle = \langle W_1, \mathcal{P}_{T^\perp} C^* \rangle = \|\mathcal{P}_{T^\perp} C^*\|_*.$$

Similarly, the subgradient of $\lambda\|C\|_1$ is $\lambda \mathrm{sgn}(C) + W_2$, for any W_2 obeying $\mathrm{supp}(W_2) \subseteq \Omega^c$ and $\|W_2\|_\infty \leq \lambda$. We may choose W_2 such that $\|W_2\|_\infty = \lambda$ and $\langle [W_2]_{\Omega^c}, C^*_{\Omega^c} \rangle = \|C^*_{\Omega^c}\|_1$, then by the convexity of one norm,

$$\lambda\|C^*\|_1 \geq \lambda\|C\|_1 + \lambda\langle \partial\|C\|_1, C^* - C \rangle = \lambda\|C\|_1 + \langle \lambda \mathrm{sgn}(C_\Omega), C^*_\Omega - C_\Omega \rangle + \lambda\|C^*_{\Omega^c}\|_1. \tag{C.3}$$

Then we may combine (C.2) and (C.3) with condition ① and ③ to get

$$\begin{aligned}
&\|C^*\|_* + \lambda\|C^*\|_1 \geq \ \|C\|_* + \langle UV^T, \mathcal{P}_T(C^* - C)\rangle + \|\mathcal{P}_{T^\perp}(C^*)\|_* + \lambda\|C\|_1 \\
&\quad + \langle \lambda \mathrm{sgn}(C_\Omega), C^*_\Omega - C_\Omega\rangle + \lambda\|C^*_{\Omega^c}\|_1 \\
&= \|C\|_* + \langle \mathcal{P}_T(A^T\Lambda_1 - \Lambda_2 - \Lambda_3), \mathcal{P}_T(C^* - C)\rangle + \|\mathcal{P}_{T^\perp}(C^*)\|_* + \lambda\|C\|_1 \\
&\quad + \langle \Lambda_2, C^*_\Omega - C_\Omega\rangle + \lambda\|C^*_{\Omega^c\cap\tilde{\Omega}}\|_1 + \lambda\|C^*_{\tilde{\Omega}^c}\|_1. \qquad \text{(C.4)}
\end{aligned}$$

By Lemma C.1, we know

$$\begin{aligned}
&\langle \mathcal{P}_T(A^T\Lambda_1 - \Lambda_2 - \Lambda_3), \mathcal{P}_T(C^* - C)\rangle \\
=&\langle A^T\Lambda_1 - \Lambda_2 - \Lambda_3, \mathcal{P}_T(\mathcal{P}_T(C^* - C))\rangle \\
=&\langle A^T\Lambda_1 - \Lambda_2 - \Lambda_3, \mathcal{P}_T(C^*)\rangle - \langle A^T\Lambda_1 - \Lambda_2 - \Lambda_3, \mathcal{P}_T(C)\rangle \\
=&\langle \Lambda_1, A\mathcal{P}_T(C^*)\rangle - \langle \Lambda_2 + \Lambda_3, \mathcal{P}_T(C^*)\rangle - \langle \Lambda_1, AC\rangle + \langle \Lambda_2 + \Lambda_3, C\rangle \\
=&\langle \Lambda_1, AC^* - AC\rangle - \langle \Lambda_1, A\mathcal{P}_{T^\perp}(C^*)\rangle + \langle \Lambda_2 + \Lambda_3, C\rangle - \langle \Lambda_2 + \Lambda_3, \mathcal{P}_T(C^*)\rangle \\
=&-\langle \Lambda_1, A\mathcal{P}_{T^\perp}(C^*)\rangle + \langle \Lambda_2 + \Lambda_3, C\rangle - \langle \Lambda_2 + \Lambda_3, C^*\rangle + \langle \Lambda_2 + \Lambda_3, \mathcal{P}_{T^\perp}(C^*)\rangle \\
=&-\langle A^T\Lambda_1 - \Lambda_2 - \Lambda_3, \mathcal{P}_{T^\perp}(C^*)\rangle - \langle \Lambda_2 + \Lambda_3, C^*\rangle + \langle \Lambda_2 + \Lambda_3, C\rangle \\
=&-\langle \mathcal{P}_{T^\perp}(A^T\Lambda_1 - \Lambda_2), \mathcal{P}_{T^\perp}(C^*)\rangle - \langle \Lambda_2 + \Lambda_3, C^*\rangle + \langle \Lambda_2 + \Lambda_3, C\rangle \\
=&-\langle \mathcal{P}_{T^\perp}(A^T\Lambda_1 - \Lambda_2), \mathcal{P}_{T^\perp}(C^*)\rangle - \langle \Lambda_2, C^*\rangle + \langle \Lambda_2, C\rangle.
\end{aligned}$$

Note that the last step follows from condition ⑥ and C, C^*'s primal feasibility. Substitute back into (C.4), we get

$$\begin{aligned}
&\|C^*\|_* + \lambda\|C^*\|_1 \\
\geq&\|C\|_* + \lambda\|C\|_1 + \|\mathcal{P}_{T^\perp}(C^*)\|_* - \langle \mathcal{P}_{T^\perp}(A^T\Lambda_1 - \Lambda_2 - \Lambda_3), \mathcal{P}_{T^\perp}(C^*)\rangle \\
&+ \lambda\|C^*_{\Omega^c\cap\tilde{\Omega}}\|_1 - \langle [\Lambda_2]_{\Omega^c\cap\tilde{\Omega}}, C^*_{\Omega^c\cap\tilde{\Omega}}\rangle + \lambda\|C^*_{\tilde{\Omega}^c}\|_1 - \langle [\Lambda_2]_{\tilde{\Omega}^c}, C^*_{\tilde{\Omega}^c}\rangle \\
\geq&\|C\|_* + \lambda\|C\|_1 - (1 - \|\mathcal{P}_{T^\perp}(A^T\Lambda_1 - \Lambda_2 - \Lambda_3)\|)\|\mathcal{P}_{T^\perp}(C^*)\|_* \\
&(\lambda - \|[\Lambda_2]_{\Omega^c\cap\tilde{\Omega}}\|_\infty)\|C^*_{\Omega^c\cap\tilde{\Omega}}\|_1 + (\lambda - \|[\Lambda_2]_{\tilde{\Omega}^c}\|_\infty)\|C^*_{\tilde{\Omega}^c}\|_1
\end{aligned}$$

Assume $C^*_{\tilde{\Omega}^c} \neq 0$. By condition ④, ⑤ and ②, we have the strict inequality

$$\|C^*\|_* + \lambda\|C^*\|_1 > \|C\|_* + \lambda\|C\|_1.$$

Recall that C^* is an optimal solution, i.e., $\|C^*\|_* + \lambda\|C^*\|_1 \leq \|C\|_* + \lambda\|C\|_1$. By contradiction, we conclude that $C^*_{\tilde{\Omega}^c} = 0$ for any optimal solution C^*. □

C.1.2 Constructing solution

Apply Lemma C.2 with $A = X$, $B = X$ and $\tilde{\Omega}$ guarantees the Self-Expressiveness Property (SEP), then if we can find Λ_1 and Λ_2 satisfying the five conditions with respect to a feasible C, then we know all optimal solutions of (4.1) obey SEP. The dimension of the dual variables are $\Lambda_1 \in \mathbb{R}^{n\times N}$ and $\Lambda_2 \in \mathbb{R}^{N\times N}$.

First layer fictitious problem

A good candidate can be constructed by the optimal solutions of the fictitious programs for $i = 1, 2, 3$

$$\mathbf{P_1}: \quad \min_{C_1^{(i)}, C_2^{(i)}} \|C_1^{(i)}\|_* + \lambda\|C_2^{(i)}\|_1 \text{s.t.} \quad X^{(i)} = XC_1^{(i)},\ C_1^{(i)} = C_2^{(i)},\ \mathcal{P}_{D_i}(C_1^{(i)}) = 0. \tag{C.5}$$

Corresponding dual problem is

$$\begin{aligned} \mathbf{D_1}: \quad & \max_{\Lambda_1^{(i)}, \Lambda_2^{(i)}, \Lambda_3^{(i)}} \langle X^{(i)}, \Lambda_1^{(i)} \rangle \\ & \text{s.t.} \quad \|\Lambda_2^{(i)}\|_\infty \leq \lambda,\ \|X^T\Lambda_1^{(i)} - \Lambda_2^{(i)} - \Lambda_3^{(i)}\| \leq 1,\ \mathcal{P}_{D_i^c}(\Lambda_3^{(i)}) = 0 \end{aligned} \tag{C.6}$$

where $\Lambda_1^{(i)} \in \mathbb{R}^{n\times N_i}$ and $\Lambda_2^{(i)}, \Lambda_3^{(i)} \in \mathbb{R}^{N\times N_i}$. D_i is the diagonal set of the i^{th} $N_i \times N_i$ block of $C_1^{(i)}$. For instance for $i = 2$,

$$C_1^{(2)} = \begin{pmatrix} 0 \\ \tilde{C}_1^{(2)} \\ 0 \end{pmatrix}, \qquad D_2 = \left\{ (i,j) \,\middle|\, \begin{bmatrix} 0 \\ I \\ 0 \end{bmatrix}_{ij} \neq 0 \right\},$$

The candidate solution is $C = \left[C_1^{(1)}\ C_1^{(2)}\ C_1^{(3)}\right]$. Now we need to use a second layer of fictitious problem and the same Lemma C.2 with $A = X$, $B = X^{(i)}$ to show that the solution support $\tilde{\Omega}^{(i)}$ is like the following

$$C_1^{(1)} = \begin{pmatrix} \tilde{C}_1^{(1)} \\ 0 \\ 0 \end{pmatrix}, \quad C_1^{(2)} = \begin{pmatrix} 0 \\ \tilde{C}_1^{(2)} \\ 0 \end{pmatrix}, \quad C_1^{(3)} = \begin{pmatrix} 0 \\ 0 \\ \tilde{C}_1^{(3)} \end{pmatrix}. \tag{C.7}$$

Second layer fictitious problem

The second level of fictitious problems are used to construct a suitable solution. Consider for $i = 1, 2, 3$,

$$\begin{aligned} \mathbf{P_2}: \quad & \min_{\tilde{C}_1^{(i)}, \tilde{C}_2^{(i)}} \ \|\tilde{C}_1^{(i)}\|_* + \lambda \|\tilde{C}_2^{(i)}\|_1 \\ & s.t. \quad X^{(i)} = X^{(i)} \tilde{C}_1^{(i)},\ \tilde{C}_1^{(i)} = \tilde{C}_2^{(i)},\ \mathrm{diag}(\tilde{C}_1^{(i)}) = 0. \end{aligned} \tag{C.8}$$

which is apparently feasible. Note that the only difference between the second layer fictitious problem (C.8) and the first layer fictitious problem (C.5) is the dictionary/design matrix being used. In (C.5), the dictionary contains all data points, whereas here in (C.8), the dictionary is nothing but $X^{(i)}$ itself. The corresponding dimension of representation matrix $C_1^{(i)}$ and $\tilde{C}_1^{(i)}$ are of course different too. Sufficiently we hope to establish the conditions where the solutions of (C.8) and (C.5) are related by (C.7).

The corresponding dual problem is

$$\begin{aligned} \mathbf{D_2}: \quad & \max_{\tilde{\Lambda}_1^{(i)}, \tilde{\Lambda}_2^{(i)}, \tilde{\Lambda}_3^{(i)}} \ \langle X^{(i)}, \tilde{\Lambda}_1^{(i)} \rangle \\ & s.t. \quad \|\tilde{\Lambda}_2^{(i)}\|_\infty \le \lambda,\ \|[X^{(i)}]^T \tilde{\Lambda}_1^{(i)} - \tilde{\Lambda}_2^{(i)} - \tilde{\Lambda}_3^{(i)}\| \le 1,\ \mathrm{diag}^\perp\left(\tilde{\Lambda}_3^{(i)}\right) = 0 \end{aligned} \tag{C.9}$$

where $\tilde{\Lambda}_1^{(i)} \in \mathbb{R}^{n \times N_i}$ and $\tilde{\Lambda}_2^{(i)}, \tilde{\Lambda}_3^{(i)} \in \mathbb{R}^{N_i \times N_i}$.

The proof is two steps. First we show the solution of (C.8), zero padded as in (C.7) are indeed optimal solutions of (C.5) and verify that all optimal solutions have such shape using Lemma C.2. The second step is to verify that solution $C = \left[C_1^{(1)}\ C_1^{(2)}\ C_1^{(3)}\right]$

is optimal solution of (4.1).

C.1.3 Constructing dual certificates

To complete the first step, we need to construct $\Lambda_1^{(i)}$, $\Lambda_2^{(i)}$ and $\Lambda_3^{(i)}$ such that all conditions in Lemma C.2 are satisfied. We use $i = 1$ to illustrate. Let the optimal solution[1] of (C.9) be $\tilde{\Lambda}_1^{(1)}$, $\tilde{\Lambda}_2^{(1)}$ and $\tilde{\Lambda}_3^{(1)}$. We set

$$\Lambda_1^{(1)} = \tilde{\Lambda}_1^{(1)} \qquad \Lambda_2^{(1)} = \begin{pmatrix} \tilde{\Lambda}_2^{(1)} \\ \Lambda_a \\ \Lambda_b \end{pmatrix} \qquad \text{and} \qquad \Lambda_3^{(1)} = \begin{pmatrix} \tilde{\Lambda}_3^{(1)} \\ 0 \\ 0 \end{pmatrix}$$

As $\tilde{\Omega}$ defines the first block now, this construction naturally guarantees ③ and ④. ⑥ follows directly from the dual feasibility. The existence of Λ_a and Λ_b obeying ⑤①② is something we need to show.

To evaluate ① and ②, let's first define the projection operator. Take skinny SVD $\tilde{C}_1^{(1)} = \tilde{U}^{(1)}\tilde{\Sigma}^{(1)}(\tilde{V}^{(1)})^T$.

$$C_1^{(1)} = \begin{pmatrix} \tilde{C}_1^{(1)} \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} \tilde{U}^{(1)} \\ 0 \\ 0 \end{pmatrix} \tilde{\Sigma}^{(1)}(\tilde{V}^{(1)})^T$$

$$U^{(1)}[U^{(1)}]^T = \begin{pmatrix} \tilde{U}^{(1)}[\tilde{U}^{(1)}]^T & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \qquad V^{(1)}[V^{(1)}]^T = \tilde{V}^{(1)}(\tilde{V}^{(1)})^T$$

[1] It need not be unique, for now we just use them to denote any optimal solution.

For condition ① we need

$$\mathcal{P}_{T_1}\left(X^T\Lambda_1^{(1)}-\Lambda_2^{(1)}\right)=\mathcal{P}_{T_1}\begin{pmatrix}[X^{(1)}]^T\tilde{\Lambda}_1^{(1)}-\tilde{\Lambda}_2-\tilde{\Lambda}_3\\ [X^{(2)}]^T\tilde{\Lambda}_1^{(1)}-\Lambda_a\\ [X^{(3)}]^T\tilde{\Lambda}_1^{(1)}-\Lambda_b\end{pmatrix}$$
$$=\begin{pmatrix}\mathcal{P}_{\tilde{T}_1}([X^{(1)}]^T\tilde{\Lambda}_1^{(1)}-\tilde{\Lambda}_2-\tilde{\Lambda}_3)\\ ([X^{(2)}]^T\tilde{\Lambda}_1^{(1)}-\Lambda_a)\tilde{V}^{(1)}(\tilde{V}^{(1)})^T\\ ([X^{(3)}]^T\tilde{\Lambda}_1^{(1)}-\Lambda_b)\tilde{V}^{(1)}(\tilde{V}^{(1)})^T\end{pmatrix}=\begin{pmatrix}\tilde{U}^{(1)}[\tilde{V}^{(1)}]^T\\ 0\\ 0\end{pmatrix}$$

The first row is guaranteed by construction. The second and third row are something we need to show. For condition ②

$$\left\|\mathcal{P}_{T_1^\perp}\left(X^T\Lambda_1^{(1)}-\Lambda_2^{(1)}-\tilde{\Lambda}_3\right)\right\|=\left\|\begin{pmatrix}\mathcal{P}_{\tilde{T}_1^\perp}([X^{(1)}]^T\tilde{\Lambda}_1^{(1)}-\tilde{\Lambda}_2-\tilde{\Lambda}_3)\\ ([X^{(2)}]^T\tilde{\Lambda}_1^{(1)}-\Lambda_a)(I-\tilde{V}^{(1)}(\tilde{V}^{(1)})^T)\\ ([X^{(3)}]^T\tilde{\Lambda}_1^{(1)}-\Lambda_b)(I-\tilde{V}^{(1)}(\tilde{V}^{(1)})^T)\end{pmatrix}\right\|$$
$$\leq\|\mathcal{P}_{\tilde{T}_1^\perp}([X^{(1)}]^T\tilde{\Lambda}_1^{(1)}-\tilde{\Lambda}_2-\tilde{\Lambda}_3)\|+\|[X^{(2)}]^T\tilde{\Lambda}_1^{(1)}-\Lambda_a\|+\|[X^{(3)}]^T\tilde{\Lambda}_1^{(1)}-\Lambda_b\|$$

Note that as $([X^{(2)}]^T\tilde{\Lambda}_1^{(1)}-\Lambda_a)\tilde{V}^{(1)}(\tilde{V}^{(1)})^T=0$, the complement projection $([X^{(2)}]^T\tilde{\Lambda}_1^{(1)}-\Lambda_a)(I-\tilde{V}^{(1)}(\tilde{V}^{(1)})^T)=([X^{(2)}]^T\tilde{\Lambda}_1^{(1)}-\Lambda_a)$. The same goes for the third row. In fact, in worst case, $\|\mathcal{P}_{\tilde{T}_1^\perp}([X^{(1)}]^T\tilde{\Lambda}_1^{(1)}-\tilde{\Lambda}_2)\|=1$, then for both ① and ② to hold, we need

$$[X^{(2)}]^T\tilde{\Lambda}_1^{(1)}-\Lambda_a=0,\qquad\qquad [X^{(3)}]^T\tilde{\Lambda}_1^{(1)}-\Lambda_b=0. \tag{C.10}$$

In other words, the conditions reduce to whether there exist Λ_a,Λ_b obeying entry-wise box constraint λ that can nullify $[X^{(2)}]^T\tilde{\Lambda}_1^{(1)}$ and $[X^{(3)}]^T\tilde{\Lambda}_1^{(1)}$.

In fact, as we will illustrate, (C.10) is sufficient for the original optimization (4.1) too. We start the argument by taking the skinny SVD of constructed solution C.

$$C=\begin{pmatrix}\tilde{C}_1&0&0\\0&\tilde{C}_2&0\\0&0&\tilde{C}_3\end{pmatrix}=\begin{pmatrix}\tilde{U}_1&0&0\\0&\tilde{U}_2&0\\0&0&\tilde{U}_3\end{pmatrix}\begin{pmatrix}\tilde{\Sigma}_1&0&0\\0&\tilde{\Sigma}_2&0\\0&0&\tilde{\Sigma}_3\end{pmatrix}\begin{pmatrix}\tilde{V}_1&0&0\\0&\tilde{V}_2&0\\0&0&\tilde{V}_3\end{pmatrix}.$$

Check that U, V are both orthonormal, Σ is diagonal matrix with unordered singular values. Let the block diagonal shape be Ω, the five conditions in Lemma C.2 are met with

$$\Lambda_1 = \begin{pmatrix} \tilde{\Lambda}_1^{(1)} & \tilde{\Lambda}_1^{(2)} & \tilde{\Lambda}_1^{(3)} \end{pmatrix}, \quad \Lambda_2 = \begin{pmatrix} \tilde{\Lambda}_2^{(1)} & \Lambda_a^{(2)} & \Lambda_a^{(3)} \\ \Lambda_a^{(1)} & \tilde{\Lambda}_2^{(2)} & \Lambda_b^{(3)} \\ \Lambda_b^{(1)} & \Lambda_b^{(2)} & \tilde{\Lambda}_2^{(3)} \end{pmatrix}, \quad \Lambda_3 = \begin{pmatrix} \tilde{\Lambda}_3^{(1)} & 0 & 0 \\ 0 & \tilde{\Lambda}_3^{(2)} & 0 \\ 0 & 0 & \tilde{\Lambda}_3^{(3)} \end{pmatrix},$$

as long as $\Lambda_1^{(i)}$, $\Lambda_2^{(i)}$ and $\Lambda_3^{(i)}$ guarantee the optimal solution of (C.5) obeys SEP for each i. Condition ③ ④ ⑤ and ⑥ are trivial. To verify condition ① and ②,

$$\begin{aligned} & X^T\Lambda_1 - \Lambda_2 - \Lambda_3 \\ = & \begin{pmatrix} [X^{(1)}]^T\tilde{\Lambda}_1^{(1)} - \tilde{\Lambda}_2^{(1)} - \tilde{\Lambda}_3^{(1)} & [X^{(1)}]^T\tilde{\Lambda}_1^{(2)} - \Lambda_a^{(2)} & [X^{(1)}]^T\tilde{\Lambda}_1^{(3)} - \Lambda_a^{(3)} \\ [X^{(2)}]^T\tilde{\Lambda}_1^{(1)} - \Lambda_a^{(1)} & [X^{(2)}]^T\tilde{\Lambda}_1^{(2)} - \tilde{\Lambda}_2^{(2)} - \tilde{\Lambda}_3^{(2)} & [X^{(2)}]^T\tilde{\Lambda}_1^{(3)} - \Lambda_b^{(3)} \\ [X^{(3)}]^T\tilde{\Lambda}_1^{(1)} - \Lambda_b^{(1)} & [X^{(3)}]^T\tilde{\Lambda}_1^{(2)} - \Lambda_b^{(2)} & [X^{(3)}]^T\tilde{\Lambda}_1^{(3)} - \tilde{\Lambda}_2^{(3)} - \tilde{\Lambda}_3^{(3)} \end{pmatrix} \\ = & \begin{pmatrix} [X^{(1)}]^T\tilde{\Lambda}_1^{(1)} - \tilde{\Lambda}_2^{(1)} - \tilde{\Lambda}_3^{(1)} & 0 & 0 \\ 0 & [X^{(2)}]^T\tilde{\Lambda}_1^{(2)} - \tilde{\Lambda}_2^{(2)} - \tilde{\Lambda}_3^{(2)} & 0 \\ 0 & 0 & [X^{(3)}]^T\tilde{\Lambda}_1^{(3)} - \tilde{\Lambda}_2^{(3)} - \tilde{\Lambda}_3^{(3)} \end{pmatrix}. \end{aligned}$$

Furthermore, by the block-diagonal SVD of C, projection $\mathcal{P}_T$ can be evaluated for each diagonal block, where optimality condition of the second layer fictitious problem guarantees that for each i

$$\mathcal{P}_{\tilde{T}_i}([X^{(i)}]^T\tilde{\Lambda}_1^{(i)} - \tilde{\Lambda}_2^{(i)} - \tilde{\Lambda}_3^{(i)}) = \tilde{U}_i\tilde{V}_i^T.$$

It therefore holds that

$$\textcircled{1} \quad \mathcal{P}_T(X^T\Lambda_1 - \Lambda_2 - \Lambda_3) = \begin{pmatrix} \tilde{U}_1\tilde{V}_1^T & 0 & 0 \\ 0 & \tilde{U}_2\tilde{V}_2^T & 0 \\ 0 & 0 & \tilde{U}_3\tilde{V}_3^T \end{pmatrix} = UV^T,$$

$$\begin{aligned} \textcircled{2} \quad & \|\mathcal{P}_{T^\perp}(X^T\Lambda_1 - \Lambda_2)\| \\ &= \left\| \begin{array}{ccc} \mathcal{P}_{\tilde{T}_i^\perp}([X^{(1)}]^T\tilde{\Lambda}_1^{(1)} - \tilde{\Lambda}_2^{(1)}) & 0 & 0 \\ 0 & \mathcal{P}_{\tilde{T}_i^\perp}([X^{(2)}]^T\tilde{\Lambda}_1^{(2)} - \tilde{\Lambda}_2^{(2)}) & 0 \\ 0 & 0 & \mathcal{P}_{\tilde{T}_i^\perp}([X^{(3)}]^T\tilde{\Lambda}_1^{(3)} - \tilde{\Lambda}_2^{(3)}) \end{array} \right\| \\ &= \max_{i=1,2,3} \|\mathcal{P}_{\tilde{T}_i^\perp}([X^{(1)}]^T\tilde{\Lambda}_1^{(i)} - \tilde{\Lambda}_2^{(i)})\| \leq 1. \end{aligned}$$

C.1.4 Dual Separation Condition

Definition C.1 (Dual Separation Condition)**.** *For $X^{(i)}$, if the corresponding dual optimal solution $\tilde{\Lambda}_1^{(i)}$ of* (C.9) *obeys $\|[X^{(j)}]^T\tilde{\Lambda}_1^{(i)}\|_\infty < \lambda$ for all $j \neq i$, then we say that dual separation condition holds.*

Remark C.1. Definition C.1 directly implies the existence of Λ_a, Λ_b obeying (C.10).

Bounding $\|[X^{(j)}]^T\tilde{\Lambda}_1^{(i)}\|_\infty$ is equivalent to bound the maximal inner product of arbitrary column pair of $X^{(j)}$ and $\tilde{\Lambda}_1^{(i)}$. Let x be a column of $X^{(j)}$ and ν be a column of $\tilde{\Lambda}_1^{(i)}$,

$$\langle x, \nu\rangle = \|\nu^*\|\langle x, \frac{\nu}{\|\nu^*\|}\rangle \leq \|\nu^*\|\|[V^{(i)}]^T x\|_\infty \leq \max_k \|\mathrm{Proj}_{\mathcal{S}_i}(\tilde{\Lambda}_1^{(i)})\mathbf{e}_k\| \max_{x\in\mathcal{X}\setminus\mathcal{X}_i} \|[V^{(i)}]^T x\|_\infty.$$

where $V^{(i)} = [\frac{\nu_1}{\|\nu_1^*\|}, ..., \frac{\nu_{N_i}}{\|\nu_{N_i}^*\|}]$ is a normalized dual matrix as defined in Definition 4.2 and $\mathbf{e}_k$ denotes standard basis. Recall that in Definition 4.2, ν^* is the component of ν inside $\mathcal{S}_i$ and ν is normalized such that $\|\nu^*\| = 1$. It is easy to verify that $[\tilde{\Lambda}_1^{(i)}]^* = \mathrm{Proj}_{\mathcal{S}_i}(\tilde{\Lambda}_1^{(i)})$ is minimum-Frobenious-norm optimal solution. Note that we can choose $\tilde{\Lambda}_1^{(i)}$ to be any optimal solution of (C.9), so we take $\tilde{\Lambda}_1^{(i)}$ such that the associated $V^{(i)}$ is the one that minimizes $\max_{x\in\mathcal{X}\setminus\mathcal{X}_i} \|[V^{(i)}]^T x\|_\infty$.

Now we may write a sufficient dual separation condition in terms of the incoherence μ in Definition 4.3,

$$\langle x, \nu \rangle \leq \max_k \|[\tilde{\Lambda}_1^{(i)}]^* \mathbf{e}_k\| \mu(\mathcal{X}_i) \leq \lambda. \tag{C.11}$$

Now it is left to bound $\max_k \|[\tilde{\Lambda}_1^{(i)}]^* \mathbf{e}_k\|$ with meaningful properties of $X^{(i)}$.

C.1.4.1 Separation condition via singular value

By the second constraint of (C.9), we have

$$1 \geq \|[X^{(i)}]^T \tilde{\Lambda}_1^{(i)} - \tilde{\Lambda}_2^{(i)} - \tilde{\Lambda}_3^{(i)}\| \geq \max_k \|([X^{(i)}]^T \tilde{\Lambda}_1^{(i)} - \tilde{\Lambda}_2^{(i)} - \tilde{\Lambda}_3^{(i)}) \mathbf{e}_k\| := \|v\| \tag{C.12}$$

Note that $\max_k \|([X^{(i)}]^T \tilde{\Lambda}_1^{(i)} - \tilde{\Lambda}_2^{(i)} - \tilde{\Lambda}_3^{(i)}) \mathbf{e}_k\|$ is the 2-norm of a vector and we conveniently denote this vector by v. It follows that

$$\|v\| = \sqrt{|v_k|^2 + \sum_{i \neq k} |v_i|^2} \geq \sqrt{\sum_{i \neq k} |v_i|^2} = \|v_{-k}\|, \tag{C.13}$$

where v_k denotes the k^{th} element and v_{-k} stands for v with the k^{th} element removed. For convenience, we also define X_{-k} to be X with the k^{th} column removed and X_k to be the k^{th} column vector of X.

By condition ⑥ in Lemma C.2, $\tilde{\Lambda}_3^{(i)}$ is diagonal, hence $\tilde{\Lambda}_3^{(i)} \mathbf{e}_k = \left[0, ..., [\tilde{\Lambda}_3^{(i)} \mathbf{e}_k]_k, ..., 0\right]^T$ and $[\tilde{\Lambda}_3^{(i)} \mathbf{e}_k]_{-k} = 0$. To be precise, we may get rid of $\tilde{\Lambda}_3^{(i)}$ all together

$$\|v_{-k}\| = \max_k \left\| \left([X_{-k}^{(i)}]^T \tilde{\Lambda}_1^{(i)} - [[\tilde{\Lambda}_2^{(i)}]^T]_{-k} \right) \mathbf{e}_k \right\|.$$

Note that $\max_k \|X \mathbf{e}_k\|$ is a norm, as is easily shown in the following lemma.

Lemma C.3. *Function $f(X) := \max_k \|X \mathbf{e}_k\|$ is a norm.*

Proof. We prove by definition of a norm.

(1) $f(aX) = \max_k \|[aX]_k\| = \max_k (|a| \|X_k\|) = \|a\| f(X)$.

(2) Assume $X \neq 0$ and $f(X) = 0$. Then for some (i, j), $X_{ij} = c \neq 0$, so $f(X) \geq |c|$

which contradicts $f(X) = 0$.

(3) Triangular inequality:

$$\begin{aligned} f(X_1 + X_2) &= \max_k(\|[X_1 + X_2]_k\|) \leq \max_k(\|[X_1]_k\| + \|[X_2]_k\|) \\ &\leq \max_{k_1}(\|[X_1]_{k_1}\|) + \max_{k_2}(\|[X_2]_{k_2}\|) = f(X_1) + f(X_2). \end{aligned}$$

□

Thus by triangular inequality,

$$\begin{aligned} \|v_{-k}\| \geq& \max_k \left\| [X_{-k}^{(i)}]^T [\tilde{\Lambda}_1^{(i)} \mathbf{e}_k] \right\| - \max_k \left\| [[\tilde{\Lambda}_2^{(i)}]^T]_{-k} \mathbf{e}_k \right\| \\ \geq& \sigma_{d_i}(X_{-k}^{(i)}) \max_k \|[\tilde{\Lambda}_1^{(i)}]^* \mathbf{e}_k\| - \lambda\sqrt{N_i - 1} \end{aligned} \tag{C.14}$$

where $\sigma_{d_i}(X_{-k}^{(i)})$ is the r^{th} (smallest non-zero) singular value of $X_{-k}^{(i)}$. The last inequality is true because $X_{-k}^{(i)}$ and $[\tilde{\Lambda}_1^{(i)}]^*$ belong to the same d_i-dimensional subspace and the condition $\|\tilde{\Lambda}_2^{(i)}\|_\infty \leq \lambda$. Combining (C.12)(C.13) and (C.14), we find the desired bound

$$\max_k \|[\tilde{\Lambda}_1^{(i)}]^* \mathbf{e}_k\| \leq \frac{1 + \lambda\sqrt{N_i - 1}}{\sigma_{d_i}(X_{-k}^{(i)})} < \frac{1 + \lambda\sqrt{N_i}}{\sigma_{d_i}(X_{-k}^{(i)})}.$$

The condition (C.11) now becomes

$$\langle x, \nu \rangle \leq \frac{\mu(1 + \lambda\sqrt{N_i})}{\sigma_{d_i}(X_{-k}^{(i)})} < \lambda \;\Leftrightarrow\; \mu(1 + \lambda\sqrt{N_i}) < \lambda\sigma_{d_i}(X_{-k}^{(i)}). \tag{C.15}$$

Note that when $X^{(i)}$ is well conditioned with condition number κ,

$$\sigma_{d_i}(X_{-k}^{(i)}) = \frac{1}{\kappa\sqrt{d_i}} \|X_{-k}^{(i)}\|_F = (1/\kappa)\sqrt{N_i/d_i}.$$

To interpret the inequality, we remark that when $\mu\kappa\sqrt{d_i} < 1$ there always exists a λ such that SEP holds.

C.1.4.2 Separation condition via inradius

This time we relax the inequality in (C.14) towards the max/infinity norm.

$$\begin{aligned}\|v_{-k}\| &= \max_k \left\| \left([X_{-k}^{(i)}]^T \tilde{\Lambda}_1^{(i)} - [[\tilde{\Lambda}_2^{(i)}]^T]_{-k} \right) \mathbf{e}_k \right\| \\ &\geq \max_k \left\| \left([X_{-k}^{(i)}]^T \tilde{\Lambda}_1^{(i)} - [[\tilde{\Lambda}_2^{(i)}]^T]_{-k} \right) \mathbf{e}_k \right\|_\infty \\ &\geq \max_k \left\| [X_{-k}^{(i)}]^T [\tilde{\Lambda}_1^{(i)}]^* \right\|_\infty - \lambda \end{aligned} \tag{C.16}$$

This is equivalent to for all $k = 1, .., N_i$

$$\begin{cases} \|[X_{-k}^{(i)}]^T \nu_1^*\|_\infty \leq 1+\lambda, \\ \|[X_{-k}^{(i)}]^T \nu_2^*\|_\infty \leq 1+\lambda, \\ ... \\ \|[X_{-k}^{(i)}]^T \nu_{N_i}^*\|_\infty \leq 1+\lambda, \end{cases} \Leftrightarrow \begin{cases} \nu_1^* \in (1+\lambda)[\mathrm{conv}(\pm X_{-k}^{(i)})]^o, \\ \nu_2^* \in (1+\lambda)[\mathrm{conv}(\pm X_{-k}^{(i)})]^o, \\ ... \\ \nu_{N_i}^* \in (1+\lambda)[\mathrm{conv}(\pm X_{-k}^{(i)})]^o, \end{cases}$$

where $\mathcal{P}^o$ represents the polar set of a convex set $\mathcal{P}$, namely, every column of $\tilde{\Lambda}_1^{(i)}$ in (C.11) is within this convex polytope $[\mathrm{conv}(\pm X_{-k}^{(i)})]^o$ scaled by $(1+\lambda)$. A upper bound follows from the geometric properties of the symmetric convex polytope.

Definition C.2 (circumradius)**.** *The circumradius of a convex body $\mathcal{P}$, denoted by $R(\mathcal{P})$, is defined as the radius of the smallest Euclidean ball containing $\mathcal{P}$.*

The magnitude $\|\nu^*\|$ is bounded by $R([\mathrm{conv}(\pm X_{-k}^{(i)})]^o)$. Moreover, by the the following lemma we may find the circumradius by analyzing the polar set of $[\mathrm{conv}(\pm X_{-k}^{(i)})]^o$ instead. By the property of polar operator, polar of a polar set gives the tightest convex envelope of original set, i.e., $(\mathcal{K}^o)^o = \mathrm{conv}(\mathcal{K})$. Since $\mathrm{conv}(\pm X_{-k}^{(i)})$ is convex in the first place, the polar set is essentially $\mathrm{conv}(\pm X_{-k}^{(i)})$.

Lemma C.4. *For a symmetric convex body $\mathcal{P}$, i.e. $\mathcal{P} = -\mathcal{P}$, inradius of $\mathcal{P}$ and circumradius of polar set of $\mathcal{P}$ satisfy:*

$$r(\mathcal{P})R(\mathcal{P}^o) = 1.$$

By this observation, we have for all $j = 1, ..., N_i$

$$\|\nu_j^*\| \leq (1+\lambda)R(\text{conv}(\pm X_{-k}^{(i)})) = \frac{1+\lambda}{r(\text{conv}(\pm X_{-k}^{(i)}))}.$$

Then the condition becomes

$$\frac{\mu(1+\lambda)}{r(\text{conv}(\pm X_{-k}^{(i)}))} < \lambda \;\Leftrightarrow\; \mu(1+\lambda) < \lambda r(\text{conv}(\pm X_{-k}^{(i)})), \tag{C.17}$$

which reduces to the condition of SSC when λ is large (if we take the μ definition in [124]).

With (C.15) and (C.17), the proof for Theorem 4.1 is complete.

C.2 Proof of Theorem 4.2 (the randomized result)

Theorem 4.2 is essentially a corollary of the deterministic results. The proof of it is no more than providing probabilistic lower bounds of smallest singular value σ (Lemma 4.1), inradius (Lemma 4.2) and upper bounds for minimax subspace incoherence μ (Lemma 4.3), then use union bound to make sure all random events happen together with high probability.

C.2.1 Smallest singular value of unit column random low-rank matrices

We prove Lemma 4.1 in this section. Assume the following mechanism of random matrix generation.

1. Generate $n \times r$ Gaussian random matrix A.
2. Generate $r \times N$ Gaussian random matrix B.
3. Generate rank-r matrix AB then normalize each column to unit vector to get X.

The proof contains three steps. First is to bound the magnitude. When n is large, each column's magnitude is bounded from below with large probability. Second we

show that if we reduce the largest magnitude column to smallest column vector, the singular values are only scaled by the same factor. Thirdly use singular value bound of A and B to show that singular value of X.

$$2\sigma_r(X) > \sigma_r(AB) > \sigma_r(A)\sigma_r(B)$$

Lemma C.5 (Magnitude of Gaussian vector)**.** *For Gaussian random vector $z \in \mathbb{R}^n$, if each entry $z_i \sim N(0, \frac{\sigma}{\sqrt{n}})$, then each column z_i satisfies:*

$$Pr((1-t)\sigma^2 \leq \|z\|^2 \leq (1+t)\sigma^2) > 1 - e^{\frac{n}{2}(\log(t+1)-t)} - e^{\frac{n}{2}(\log(1-t)+t)}$$

Proof. To show the property, we observe that the sum of n independent square Gaussian random variables follows χ^2 distribution with d.o.f n, in other word, we have

$$\|z\|^2 = |z_1|^2 + ... + |z_n|^2 \sim \frac{\sigma^2}{n}\chi^2(n).$$

By Hoeffding's inequality, we have a close upper bound of its CDF [44], which gives us

$$\begin{aligned} Pr(\|z\|^2 > \alpha\sigma^2) &= 1 - \mathrm{CDF}_{\chi^2_n}(\alpha) \leq (\alpha e^{1-\alpha})^{\frac{n}{2}} && \text{for } \alpha > 1, \\ Pr(\|z\|^2 < \beta\sigma^2) &= \mathrm{CDF}_{\chi^2_n}(\beta) \leq (\beta e^{1-\beta})^{\frac{n}{2}} && \text{for } \beta < 1. \end{aligned}$$

Substitute $\alpha = 1 + t$ and $\beta = 1 - t$, and apply union bound we get exactly the concentration statement. □

To get an idea of the scale, when $t = 1/3$, the ratio of maximum and minimum $\|z\|$ is smaller than 2 with probability larger than $1 - 2\exp(-n/20)$. This proves the first step.

By random matrix theory [e.g., 45, 116, 121] asserts that G is close to an orthonormal matrix, as the following lemma, adapted from Theorem II.13 of [45], shows:

Lemma C.6 (Smallest singular value of random rectangular matrix)**.** *Let $G \in \mathbb{R}^{n\times r}$*

has i.i.d. entries $\sim N(0, 1/\sqrt{n})$*. With probability of at least* $1 - 2\gamma$*,*

$$1 - \sqrt{\frac{r}{n}} - \sqrt{\frac{2\log(1/\gamma)}{n}} \leq \sigma_{min}(G) \leq \sigma_{max}(G) \leq 1 + \sqrt{\frac{r}{n}} + \sqrt{\frac{2\log(1/\gamma)}{n}}.$$

Lemma C.7 (Smallest singular value of random low-rank matrix)**.** *Let* $A \in R^{n\times r}$*,* $B \in R^{r\times N}$*,* $r < N < n$*, furthermore,* $A_{ij} \sim N(0, 1/\sqrt{n})$ *and* $B_{ij} \sim N(0, 1/\sqrt{N})$*. Then there exists an absolute constant* C *such that with probability of at least* $1 - n^{-10}$*,*

$$\sigma_r(AB) \geq 1 - 3\sqrt{\frac{r}{N}} - C\sqrt{\frac{\log N_\ell}{N}}.$$

The proof is by simply by $\sigma_r(AB) \geq \sigma_r(A)\sigma_r(B)$, apply Lemma C.5 to both terms and then take $\gamma = \frac{1}{2N_\ell^{10}}$.

Now we may rescale each column of AB to the maximum magnitude and get $\overline{AB}$. Naturally,

$$\sigma_r(\overline{AB}) \geq \sigma_r(AB).$$

On the other hand, by the results of Step 1,

$$\sigma_r(X) \geq \sigma_r(\underline{AB}) \geq \frac{1}{2}\sigma_r(\overline{AB}) \geq \frac{1}{2}\sigma_r(AB).$$

Normalizing the scale of the random matrix and plug in the above arguments, we get Lemma 4.1 in Chapter C.

C.2.2 Smallest inradius of random polytopes

This bound in Lemma 4.2 is due to Alonso-Gutiérrez in his proof of lower bound of the volume of a random polytope[2, Lemma 3.1]. The results was made clear in the subspace clustering context by Soltanokotabi and Candes[124, Lemma 7.4]. We refer the readers to the references for the proof.

C.2.3 Upper bound of Minimax Subspace Incoherence

The upper bound of the minimax subspace incoherence (Lemma 4.3) we used in this chapter is the same as the upper bound of the subspace incoherence in [124]. This is because for by taking $V = V^*$, the value will be larger by the minimax definition[1]. For completeness, we include the steps of proof here.

The argument critically relies on the following lemma on the area of spherical cap in [6].

Lemma C.8 (Upper bound on the area of spherical cap)**.** *Let $a \in \mathbb{R}^n$ be a random vector sampled from a unit sphere and z is a fixed vector. Then we have:*

$$Pr\left(|a^T z| > \epsilon \|z\|\right) \leq 2e^{\frac{-n\epsilon^2}{2}}$$

With this result, Lemma 4.3 is proven in two steps. The first step is to apply Lemma C.8 to bound $\langle \nu_i^*, x \rangle$ and every data point $x \notin X^{(\ell)}$, where ν_i^* (a fixed vector) is the central dual vector corresponding to the data point $x_i \in X^{(\ell)}$ (see the Definition 4.3). When $\epsilon = \sqrt{\frac{6\log(N)}{n}}$, the failure probability for one even is $\frac{2}{N^3}$. Recall that ν_i^* . The second step is to use union bound across all x and then all ν_i^*. The total number of events is less than N^2 so we get

$$\mu < \sqrt{\frac{6\log N}{n}} \quad \text{with probability larger than } 1 - \frac{2}{N}.$$

C.2.4 Bound of minimax subspace incoherence for semi-random model

Another bound of the subspace incoherence can be stated under the semi-random model in [124], where subspaces are deterministic and data in each subspaces are randomly sampled. The upper bound is given as a log term times the average cosine of the canonical angles between a pair of subspaces. This is not used in this chapter, but the case of overlapping subspaces can be intuitively seen from the bound. The full statement is

[1] We did provide proof for some cases where incoherence following our new definition is significantly smaller.

rather complex and is the same form as equation (7.6) of [124], so we refer the readers there for the full proof there and only include what is different from there: the proof that central dual vector ν_i^* distributes uniformly on the unit sphere of $\mathcal{S}_\ell$.

Let U be a set of orthonormal basis of $\mathcal{S}_\ell$. Define rotation $R_{\mathcal{S}_\ell} := URU^T$ with arbitrary $d \times d$ rotation matrix R. If Λ^* be the central optimal solution of (C.9), denoted by $\mathrm{OptVal}(X^{(\ell)})$, it is easy to see that

$$R_{\mathcal{S}_\ell}\Lambda^* = \mathrm{OptVal}(R_{\mathcal{S}_\ell}X^{(\ell)}).$$

Since $X^{(\ell)}$ distribute uniformly, the probability density of getting any $X^{(\ell)}$ is identical. For each fixed instance of $X^{(\ell)}$, consider R a random variable, then the probability density of each column of Λ^* be transformed to any direction is the same. Integrating the density over all different $X^{(\ell)}$, we completed the proof for the claim that the overall probability density of ν_i^* (each column of Λ^*) pointing towards any directions in $\mathcal{S}_\ell$ is the same.

Referring to [124], the upper bound is just a concentration bound saying that the smallest inner product is close to the average cosines of the canonical angles between two subspaces, which follows from the uniform distribution of ν_i^* and uniform distribution of x in other subspaces. Therefore, when the dimension of each subspace is large, the average can still be small even though a small portion of the two subspaces are overlapping (a few canonical angles being equal to 1).

C.3 Numerical algorithm

Like described in the main text, we will derive Alternating Direction Method of Multipliers (ADMM)[17] algorithm to solve LRSSC and NoisyLRSSC. We start from noiseless version then look at the noisy version.

C.3.1 ADMM for LRSSC

First we need to reformulate the optimization with two auxiliary terms, $C = C_1 = C_2$ as in the proof to separate the two norms, and J to ensure each step has closed-form solution.

$$\min_{C_1,C_2,J} \|C_1\|_* + \lambda\|C_2\|_1 \quad s.t. \quad X = XJ, \quad J = C_2 - \mathrm{diag}(C_2), \quad J = C_1 \tag{C.18}$$

The Augmented Lagrangian is:

$$\begin{aligned}\mathcal{L} =& \|C_1\|_* + \lambda\|C_2\|_1 + \frac{\mu_1}{2}\|X - XJ\|_F^2 + \frac{\mu_2}{2}\|J - C_2 + \mathrm{diag}(C_2)\|_F^2 + \frac{\mu_3}{2}\|J - C_1\|_F^2 \\ &+ tr(\Lambda_1^T(X - XJ)) + tr(\Lambda_2^T(J - C_2 + \mathrm{diag}(C_2))) + tr(\Lambda_3^T(J - C_1)),\end{aligned}$$

where μ_1, μ_2 and μ_3 are numerical parameters to be tuned. By assigning the partial gradient/subgradient of J, C_2 and C_1 iteratively and update dual variables $\Lambda_1, \Lambda_2, \Lambda_3$ in every iterations, we obtain the update steps of ADMM.

$$J = \left[\mu_1 X^T X + (\mu_2 + \mu_3)I\right]^{-1}\left[\mu_1 X^T X + \mu_2 C_2 + \mu_3 C_1 + X^T\Lambda_1 - \Lambda_2 - \Lambda_3\right] \tag{C.19}$$

Define soft-thresholding operator $\pi_\beta(X) = (|X| - \beta)_+ \mathrm{sgn}(X)$ and singular value soft-thresholding operator $\Pi_\beta(X) = U\pi_\beta(\Sigma)V^T$, where $U\Sigma V^T$ is the skinny SVD of X. The update steps for C_1 and C_2 followed:

$$C_2 = \pi_{\frac{\lambda}{\mu_2}}\left(J + \frac{\Lambda_2}{\mu_2}\right), \quad C_2 = C_2 - \mathrm{diag}(C_2), \quad C_1 = \Pi_{\frac{1}{\mu_3}}\left(J + \frac{\Lambda_3}{\mu_3}\right). \tag{C.20}$$

Lastly, the dual variables are updated using gradient ascend:

$$\Lambda_1 = \Lambda_1 + \mu_1(X - XJ), \quad \Lambda_2 = \Lambda_2 + \mu_2(J - C_2), \quad \Lambda_3 = \Lambda_3 + \mu_3(J - C_1). \tag{C.21}$$

Algorithm 7 ADMM-LRSSC (with optional Adaptive Penalty)

Input: Data points as columns in $X \in \mathbb{R}^{n\times N}$, tradeoff parameter λ, numerical parameters $\mu_1^{(0)}, \mu_2^{(0)}, \mu_3^{(0)}$ and (optional ρ_0, μ_{max}, η,ϵ).
Initialize $C_1 = 0$, $C_2 = 0$, $J = 0$, $\Lambda_1 = 0$, $\Lambda_2 = 0$ and $\Lambda_3 = 0$.
Pre-compute X^TX and $H = \left[\mu_1 X^TX + (\mu_2+\mu_3)I\right]^{-1}$ for later use.
while not converged **do**

1. Update J by (C.19).
2. Update C_1, C_2 by (C.20).
3. Update $\Lambda_1, \Lambda_2, \Lambda_3$ by (C.21).
4. (Optional) Update parameter $(\mu_1, \mu_2, \mu_3) = \rho(\mu_1, \mu_2, \mu_3)$ and the pre-computed $H = H/\rho$ where

$$\rho = \begin{cases} \min(\mu_{max}/\mu_1, \rho_0), & \text{if } \mu_1^{\text{prev}} \max(\sqrt{\eta}\|C_1 - C_1^{\text{prev}}\|_F)/\|X\|_F \leq \epsilon; \\ 1, & \text{otherwise.} \end{cases}$$

end while
Output: Affinity matrix $W = |C_1| + |C_1|^T$

The full steps are summarized in Algorithm 7, with an optional adaptive penalty step proposed by Lin et. al[94]. Note that we deliberately constrain the proportion of μ_1, μ_2 and μ_3 such that the $\left[\mu_1 X^TX + (\mu_2+\mu_3)I\right]^{-1}$ need to be computed only once at the beginning.

C.3.2 ADMM for NoisyLRSSC

The ADMM version of NoisyLRSSC is very similar to Algorithm 7 in terms of its Lagrangian and update rule. Again, we introduce dummy variable C_1, C_2 and J to form

$$\begin{aligned} &\min_{C_1,C_2,J} \ \|X - XJ\|_F^2 + \beta_1\|C_1\|_* + \beta_2\|C_2\|_1 \\ &s.t. \quad J = C_2 - \text{diag}(C_2), \quad J = C_1. \end{aligned} \tag{C.22}$$

Its Augmented Lagrangian is

$$\begin{aligned} \mathcal{L} =& \|C_1\|_* + \lambda\|C_2\|_1 + \frac{1}{2}\|X - XJ\|_F^2 + \frac{\mu_2}{2}\|J - C_2 + \text{diag}(C_2)\|_F^2 \\ &+ \frac{\mu_3}{2}\|J - C_1\|_F^2 + tr(\Lambda_2^T(J - C_2 + \text{diag}(C_2))) + tr(\Lambda_3^T(J - C_1)), \end{aligned}$$

and update rules are:

$$J = \left[X^T X + (\mu_2 + \mu_3) I\right]^{-1} \left[X^T X + \mu_2 C_2 + \mu_3 C_1 - \Lambda_2 - \Lambda_3\right] \tag{C.23}$$

$$C_2 = \pi_{\frac{\beta_2}{\mu_2}}\left(J + \frac{\Lambda_2}{\mu_2}\right), \quad C_2 = C_2 - \mathrm{diag}(C_2), \quad C_1 = \Pi_{\frac{\beta_1}{\mu_3}}\left(J + \frac{\Lambda_3}{\mu_3}\right). \tag{C.24}$$

Update rules for Λ_2 and Λ_3 are the same as in (C.21). Note that the adaptive penalty scheme also works for NoisyLRSSC but as there is a fixed parameter in front of $X^T X$ in (C.23) now, we will need to recompute the matrix inversion every time μ_2, μ_3 get updated.

C.3.3 Convergence guarantee

Note that the general ADMM form is

$$\min_{x,z} \; f(x) + g(z) \quad s.t. \quad Ax + Bz = c. \tag{C.25}$$

In our case, $x = J$, $z = [C_1, C_2]$, $f(x) = \frac{1}{2}\|X - XJ\|_F^2$, $g(z) = \beta_1\|C_1\|_* + \beta_2\|C_2\|_1$ and constraints can be combined into a single linear equation after vectorizing J and $[C_1, C_2]$. Verify that $f(x)$ and $g(z)$ are both closed, proper and convex and the unaugmented Lagrangian has a saddle point, then the convergence guarantee follows directly from Section 3.2 in [17].

Note that the reason we can group C_1 and C_2 is because the update steps of C_1 and C_2 are concurrent and do not depends on each other (see (C.20) and (C.24) and verify). This trick is important as the convergence guarantee of the three-variable alternating direction method is still an open question.

C.4 Proof of other technical results

C.4.1 Proof of Example 4.2 (Random except 1)

Recall that the setup is L disjoint 1-dimensional subspaces in $\mathbb{R}^n$ ($L > n$). $\mathcal{S}_1, ..., \mathcal{S}_{L-1}$ subspaces are randomly drawn. $\mathcal{S}_L$ is chosen such that its angle to one of the $L-1$ subspace, say $\mathcal{S}_1$, is $\pi/6$. There is at least one samples in each subspace, so $N \geq L$. Our claim is that

Proposition C.1. *Assume the above problem setup and Definition 4.3, then with probability at least* $1 - 2L/N^3$

$$\mu \leq 2\sqrt{\frac{6\log(L)}{n}}.$$

Proof. The proof is simple. For $x_i \in \mathcal{S}_\ell$ with $\ell = 2, ..., L-1$, we simply choose $\nu_i = \nu_i^*$. Note that ν_i^* is uniformly distributed, so by Lemma C.8 and union bound, the maximum of $|\langle x, \nu_i \rangle|$ is upper bounded by $2\sqrt{\frac{6\log(N)}{n}}$ with probability at least $1 - \frac{2(L-2)^2}{N^12}$. Then we only need to consider ν_i in $\mathcal{S}_1$ and $\mathcal{S}_L$, denoted by ν_1 and ν_L. We may randomly choose any $\nu_1 = \nu_1^* + \nu_1^\perp$ obeying $\nu_1 \perp \mathcal{S}_L$ and similarly $\nu_L \perp \mathcal{S}_1$.

By the assumption that $\angle(\mathcal{S}_1, \mathcal{S}_L) = \pi/6$,

$$\|\nu_1\| = \|\nu_L\| = \frac{1}{\sin(\pi/6)} = 2.$$

Also note that they are considered a fixed vector w.r.t. all random data samples in $\mathcal{S}_2, .., \mathcal{S}_L$, so the maximum inner product is $2\sqrt{\frac{6\log(N)}{n}}$, summing up the failure probability for the remaining $2L-2$ cases, we get

$$\mu \leq 2\sqrt{\frac{6\log(N)}{n}} \quad \text{with probability } 1 - \frac{2L-2}{N^3} - \frac{2(L-2)^2}{N^{12}} > 1 - \frac{2L}{N^3}.$$

□

C.4.2 Proof of Proposition 4.1 (LRR is dense)

For easy reference, we copy the statement of Proposition 4.1 here.

Proposition C.2. *When the subspaces are independent and X is not full rank and the data points are randomly sampled from a unit sphere in each subspace, then the solution to LRR is class-wise dense, namely each diagonal block of the matrix C is all non-zero.*

Proof. The proof is of two steps. First we prove that because the data samples are random, the shape interaction matrix VV^T in Lemma 4.4 is a random projection to a rank-d_ℓ subspace in $\mathbb{R}^{N_\ell}$. Furthermore, each column is of a random direction in the subspace.

Second, we show that with probability 1, the standard bases are not orthogonal to these N_ℓ vectors inside the random subspace. The claim that VV^T is dense can hence be deduced by observing that each entry is the inner product of a column or row[1] of VV^T and a standard basis, which follows a continuous distribution. Therefore, the probability that any entries of VV^T being exactly zero is negligible. □

C.4.3 Condition (4.2) in Theorem 4.1 is computational tractable

First note that $\mu(X^{(\ell)})$ can be computed by definition, which involves solving one quadratically constrained linear program (to get dual direction matrix $[V^{(\ell)}]^*$) then finding $\mu(X^{(\ell)})$ by solving the following linear program for each subspace

$$\min_{V^{(\ell)}} \|[V^{(\ell)}]^T \overline{X^{(\ell)}}\|_\infty \quad s.t. \quad \text{Proj}_{\mathcal{S}_\ell} V^{(\ell)} = [V^{(\ell)}]^*,$$

where we use $\overline{X^{(\ell)}}$ to denote $[X^{(1)}, ..., X^{(\ell-1)}, X^{(\ell+1)}, ..., X^{(L)}]$.

To compute $\sigma_{d_\ell}(X^{(\ell)}_{-k})$, one needs to compute N_ℓ SVD of the $n \times (N_\ell - 1)$ matrix. The complexity can be further reduced by computing a close approximation of $\sigma_{d_\ell}(X^{(\ell)}_{-k})$. This can be done by finding the singular values of $X^{(\ell)}$ and use the following inequality

$$\sigma_{d_\ell}(X^{(\ell)}_{-k}) \geq \sigma_{d_\ell}(X^{(\ell)}) - 1.$$

This is a direct consequence of the SVD perturbation theory [129, Theorem 1].

[1] It makes no difference because VV^T is a symmetric matrix

C.5 Table of Symbols and Notations

Table C.1: Summary of Symbols and Notations

$\lvert\cdot\rvert$	Either absolute value or cardinality.
$\lVert\cdot\rVert$	2-norm of vector/spectral norm of matrix.
$\lVert\cdot\rVert_1$	1-norm of a vector or vectorized matrix.
$\lVert\cdot\rVert_*$	Nuclear norm/Trace norm of a matrix.
$\lVert\cdot\rVert_F$	Frobenious norm of a matrix.
$\mathcal{S}_\ell$ for $\ell = 1, .., L$	The L subspaces of interest.
n,d_ℓ	Ambient dimension, dimension of $\mathcal{S}_\ell$.
$X^{(\ell)}$	$n \times N_\ell$ matrix collecting all points from $\mathcal{S}_\ell$.
X	$n \times N$ data matrix, containing all $X^{(\ell)}$.
C	$N \times N$ Representation matrix $X = XC$. In some context, it may also denote an absolute constant.
λ	Tradeoff parameter betwenn 1-norm and nuclear norm.
A, B	Generic notation of some matrix.
$\Lambda_1, \Lambda_2, \Lambda_3$	Dual variables corresponding to the three constraints in (C.1).
$\nu, \nu_i, \nu_i^{(\ell)}$	Columns of a dual matrix.
Λ^*, ν_i^*	Central dual variables defined in Definition 4.2.
$V(X), \{V(X)\}$	Normalized dual direction matrix, and the set of all $V(X)$ (Definition 4.2).
$V^{(\ell)}$	An instance of normalized dual direction matrix $V(X^{(\ell)})$.
$v_i, v_i^{(\ell)}$	Volumns of the dual direction matrices
$\mu, \mu(X^{(\ell)})$	Incoherence parameters in Definition 4.3
σ_d, $\sigma_d(A)$	d^{th} singular value (of a matrix A).
$X_{-k}^{(\ell)}$	$X^{(\ell)}$ with k^{th} column removed.
$r, r(\mathrm{conv}(\pm X_{-k}^{(\ell)}))$	Inradius (of the symmetric convex hull of $X_{-k}^{(\ell)}$).
$\mathrm{RelViolation}(C, \mathcal{M})$	A soft measure of SEP/inter-class separation.
$\mathrm{GiniIndex}(\mathrm{vec}(C_{\mathcal{M}}))$	A soft measure of sparsity/intra-class connectivity.
Ω, $\tilde{\Omega}$, $\mathcal{M}$,$\mathcal{D}$	Some set of indices (i, j) in their respective context.
U, Σ, V	Usually the compact SVD of a matrix, e.g., C.

Continued on next page

$C_1^{(\ell)}, C_2^{(\ell)}$	Primal variables in the first layer fictitious problem.
$\tilde{C}_1^{(\ell)}, \tilde{C}_2^{(\ell)}$	Primal variables in the second layer fictitious problem.
$\Lambda_1^{(\ell)}, \Lambda_2^{(\ell)}, \Lambda_3^{(\ell)}$	Dual variables in the first layer fictitious problem.
$\tilde{\Lambda}_1^{(\ell)}, \tilde{\Lambda}_2^{(\ell)}, \tilde{\Lambda}_3^{(\ell)}$	Dual variables in the second layer fictitious problem.
$U^{(\ell)}, \Sigma^{(\ell)}, V^{(\ell)}$	Compact SVD of $C^{(\ell)}$.
$\tilde{U}^{(\ell)}, \tilde{\Sigma}^{(\ell)}, \tilde{V}^{(\ell)}$	Compact SVD of $\tilde{C}^{(\ell)}$.
$\mathrm{diag}(\cdot)$/$\mathrm{diag}^{\perp}(\cdot)$	Selection of diagonal/off-diagonal elements.
$\mathrm{supp}(\cdot)$	Support of a matrix.
$\mathrm{sgn}(\cdot)$	Sign operator on a matrix.
$\mathrm{conv}(\cdot)$	Convex hull operator.
$(\cdot)^o$	Polar operator that takes in a set and output its polar set.
$\mathrm{span}(\cdot)$	Span of a set of vectors or matrix columns.
$\mathrm{null}(\cdot)$	Nullspace of a matrix.
$\mathcal{P}_T$/$\mathcal{P}_{T^\perp}$	Projection to both column and row space of a low-rank matrix / Projection to its complement.
$\mathcal{P}_{\mathcal{D}}$	Projection to index set $\mathcal{D}$.
$\mathrm{Proj}_{\mathcal{S}}(\cdot)$	Projection to subspace $\mathcal{S}$.
β_1, β_2	Tradeoff parameters for NoisyLRSSC.
μ_1, μ_2, μ_3	Numerical parameters for the ADMM algorithm.
J	Dummy variable to formulate ADMM.

APPENDICES FOR CHAPTER 4

Appendix D

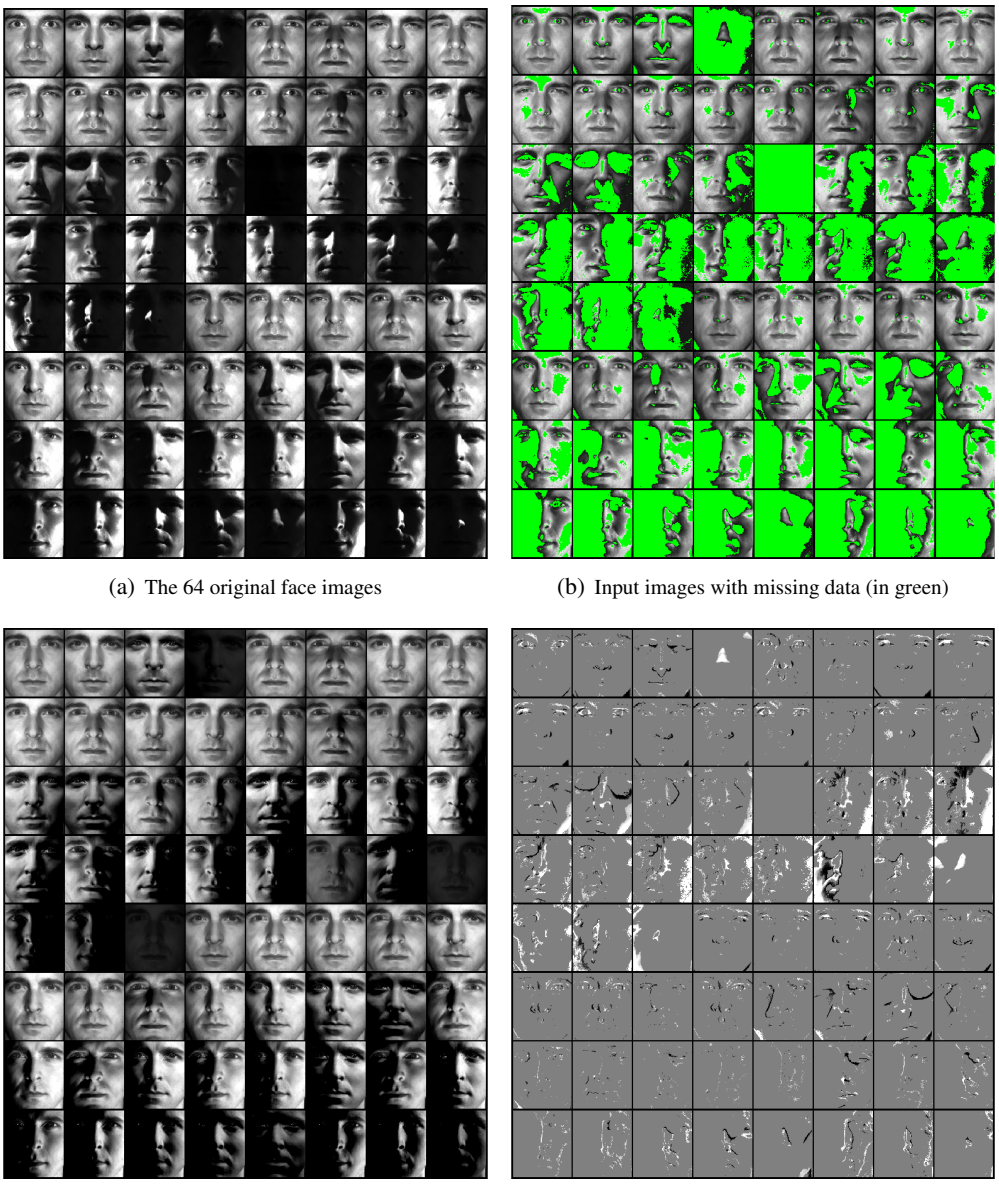
Appendices for Chapter 5

D.1 Software and source code

The point cloud in Fig. 5.16 are generated using VincentSfMToolbox [113]. Source codes of BALM, GROUSE, GRASTA, Damped Newton, Wiberg, LM_X used in the experiments are released by the corresponding author(s) of [46][7][71][19][108] and [36][1]. For Wiberg ℓ_1 [58], we have optimized the computation for Jacobian and adopted the commercial LP solver: cplex. The optimized code performs identically to the released code in small scale problems, but it is beyond the scope for us to verify for larger scale problems. In addition, we implemented SimonFunk's SVD ourselves. The ALS implementation is given in the released code package of LM_X. For OptManifold, TFOCS and CVX, we use the generic optimization packages released by the author(s) of [147][10][65] and customize for the particular problem. For NLCG, we implement the derivations in [127] and used the generic NLCG package[110].

D.2 Additional experimental results

[1]For most of these software packages, we used the default parameter in the code, or suggested by the respective authors. More careful tuning of their parameters will almost certainly result in better performances.



(a) The 64 original face images

(b) Input images with missing data (in green)

(c) The 64 recovered rank-3 face images

(d) Sparse corruptions detected

Figure D.1: Results of PARSuMi on Subject 10 of Extended YaleB. Note that the facial expressions are slightly different and some images have more than 90% of missing data. Also note that the sparse corruptions detected unified the irregular facial expressions and recovered and recovered those highlight and shadow that could not be labeled as missing data by plain thresholding.