A GENERAL FRAMEWORK FOR STRUCTURE DECOMPOSITION IN HIGH-DIMENSIONAL PROBLEMS

YANG JING

(B.Sc., SJTU, China)

A THESIS SUBMITTED
FOR THE DEGREE OF MASTER OF SCIENCE
DEPARTMENT OF MATHEMATICS
NATIONAL UNIVERSITY OF SINGAPORE
2014

To my parents

Acknowledgements

First and foremost, I would like to express my sincerest thanks to Professor Sun Defeng and Professor Xu Huan, my research supervisors, for their professional guidance and generous support throughout my time as their student.

Professor Sun, my main supervisor, has provided me with an excellent atmosphere for my study and research in optimization. His invaluable advice and insightful questions inspire me a lot. And I am deeply impressed by his immense knowledge, rigorous attitude and strong sense of responsibility in academic research. Most importantly, his encouragements have been a great support, inspiring me to go ahead when times were tough. I am truly grateful for his concern about my personal and academic development.

As my second supervisor, Professor Xu has been very supportive and has offered valuable and constructive suggestions in the way of the planning and development of this research work. I have benefited a lot from his patient guidance, scientific advice and insightful feedbacks. Without his support and active participation in the process, this thesis may never have been completed. I would like to express my very great appreciation to Professor Xu for his willingness to give his time so generously.

Thank all the fellow in the optimization research group for the informative presentations and stimulating discussions. In particular, I would like to thank Wu Bin for his guidance in research of matrix decomposition. And his assistance in thesis writing is greatly appreciated. My sincere thanks go to Yang Liuqin for his generous support and advice in coding. Many thanks to Cui Ying for her patient illustrations and sincere suggestions in studying optimization. Additionally, I would like to offer my special thanks to Yang Chen and Yu Jinjiong, who are always willing to help. Their continuous support and patient assistance are greatly appreciated.

Moreover, I would like to convey my great appreciation to National University of Singapore for the financial support and the excellent research conditions.

Last but not least, I would like to thank my family for their unconditional support throughout my degree.

Yang Jing
August 2014

Contents

Summary

In this paper, we aim to decompose the mixed structures in high-dimensional problems. We provide a general model which involves two distinct structures:

$$Y = X_1\Theta^* + X_2G^* + W,$$

where $\Theta^* \in \mathbb{R}^{p\times q}$ and $G^* \in \mathbb{R}^{n\times q}$ are some low-dimensional structured matrices, and $W \in \mathbb{R}^{n\times q}$ is the noise matrix whose Frobenius norm is assumed to be small. Then we formulate the model into the regularized squares problem and establish the M-estimator:

$$\begin{aligned}(\hat{\Theta}, \hat{G}) &\in \arg\min_{\Theta,G}\{\mathcal{L}(\Theta, G) + \lambda_1\mathcal{R}_1(\Theta) + \lambda_2\mathcal{R}_2(G)\}\\ &= \arg\min_{\Theta,G}\{\|Y - X_1\Theta - X_2G\|_F^2 + \lambda_1\mathcal{R}_1(\Theta) + \lambda_2\mathcal{R}_2(G)\},\end{aligned}$$

where $\mathcal{R}_1$ and $\mathcal{R}_2$ stand for the regularizers according to the assumed low-dimensional structures of $\Theta^* \in \mathbb{R}^{p\times q}$ and $G^* \in \mathbb{R}^{n\times q}$.

We impose four natural assumptions on the loss function $\mathcal{L}(\Theta, G)$ and the regularizers $\mathcal{R}_1(\Theta)$ and $\mathcal{R}_2(G)$ in the M-estimator. The first two basic assumptions require that the loss function is convex and differentiable, and that the regularizers are norms and decomposable. Moreover, we impose the restricted strong convexity on the loss function and require the structural incoherence property on the interaction between different structured components. Based on the four assumptions, we provide an estimation of error bound in the general model setting, which depends on the subspace compatibility constant [45] Ψ, the tuning parameters λ_1, λ_2, and some parameters in the assumed conditions.

After that, we investigate the four conditions, particularly the requirement on the structural incoherence. We then discuss the structural incoherence for different specific problems such as the PCA model. Finally, we conclude the thesis via simulations and interpret its correspondence with theoretical analysis.

List of Notations

- For any matrix A, we use A_j to denote the jth column of A.
- For any matrix A, we denote by $a_{i,j}$ the (i,j)-th entry of A.
- We use a superscript $'T'$ to represent the transpose of a matrix, i.e. A^T stands for the transpose of matrix A.
- For any matrix A, we use $\sigma_{\max}(A)$ to denote its largest singular value.
- For any random matrix X , we use X' to denote an independent copy of X.
- For each matrix A, we use $\|A\|_F$ to denote the Frobenius norm of matrix A. $\|A\|_F = \|A\|_2 = (\sum_j \sum_i |a_{i,j}|^2)^{1/2}, = \sqrt{trace(A^*A)}$
- For each matrix A, we use $\|A\|_*$ to denote its nuclear norm, i.e., the sum of the singular. values of A.
- For each matrix A, we use $\|A\|_{1,2}$ to denote the ℓ_1/ℓ_2-norm of A. $\|A\|_{1,2} = \sum_i(\sum_j |a_{i,j}|^2)^{1/2}$
- For each 1-dimensional number a, we use $|a|$ to denote the absolute value of a.

- For given subspace $\mathcal{M}$ and vector θ, we use $\Pi_{\mathcal{M}}(\theta)$ to denote the projection of vector θ onto the subspace $\mathcal{M}$.

- For a subspace pair $(\mathcal{M}, \overline{\mathcal{M}}^{\perp})$, where $\mathcal{M} \in \overline{\mathcal{M}}$, $\mathcal{M}$ represents the model subspace that captures the constraints imposed on the model parameter and is typically low-dimensional, and $\overline{\mathcal{M}}^{\perp}$ means the perturbation subspace of parameters that represents perturbations away from the model subspace.

- The subspace compatibility constant with respect to the pair $(\mathcal{R}, \|.\|)$ is defined as $\Psi(\mathcal{M}) := sup_{u \in \mathcal{M}\{0\}} \frac{\mathcal{R}(u)}{\|u\|}$.

- We use $\otimes$ to denote the cross product of two vectors or matrices.

- We use $\mathbb{E}$ to denote the operator of taking expectation.

- $\mathbb{E}_{\epsilon}$ represents expectation operator conditioning on ϵ.

- For random variable Z, we use $Var(Z)$ to denote the variance of Z.

- When we say a Σ-Gaussian random matrix B, Σ represents the second moment of B.

- We use $\mathbb{P}$ to denote the probability operator.

- We use $\mathcal{P}$ to denote the projection operator matrix.

- We use S^{p-1} to denote a sphere in p-dimensional Euclidean space.

All further notations are either standard, or defined in the text.

Chapter 1

Introduction

In many fields of science and engineering, high-dimensional problems arise in a variety of applications, including analysis of gene array data, medical imaging, remote sensing and astronomical data analysis. High-dimensional statistical inference deals with high-dimensional models in which the ambient dimension of the problem is either comparable to or possibly larger than the sample size. Since it is usually impossible to obtain consistent estimators without imposing additional model restrictions, many researchers have studied different types of structural constraints on the model (such as sparse constraints, block-wise sparse constraints, the low-rank structure and their combinations) and analyzed the behavior of the corresponding estimators. In fact, these low-dimensional constraints are motivated by structures arising in different problems.

1.1 Regularized M-estimators

For those high-dimensional problems, a general approach is to solve a regularized optimization problem which is the sum of a loss function measuring how well the model fits the data and some weighted regularization function that encourages the assumed structure. These regularized convex programs are well-known as regularized M-estimators.

Single-structured regularizers include the ℓ_1-norm regularizer for sparse constraints, the nuclear norm regularizer for low-rank requirement and the ℓ_1/ℓ_q-norm regularizer for models with block sparse structure.

For models with sparse constraints [30, 11], the ℓ_1-norm regularized estimators such as LASSO [53] or basis pursuit [19], encourage sparsity and involve solving a convex optimization problem of minimizing some variable's ℓ_1-norm. The LASSO proposed by Tibshirani [53] has gained popularity since it produces a sparse model while keeping high prediction accuracy.

Since sparsity sometimes arises in a structured manner, a line of research introduced the concept of block-wise sparse regularization [5, 61, 65, 47] and established the ℓ_1/ℓ_q-norm regularizer which is a sum (i.e., ℓ_1-norm) of ℓ_q-norms on certain subset of variables. The best known examples of such block-wise norms are the ℓ_1/ℓ_∞-norm [54, 62, 46, 63] and the ℓ_1/ℓ_2-norm [47, 38, 22]. In particular, the grouped LASSO [61] is a block-wise estimator with ℓ_1/ℓ_2-norm regularization.

For low-rank matrix estimation problems, researchers make use of the nuclear norm regularizer [15, 36, 51, 50], since it encourages sparsity in the vector of singular values and thus leads to low-rank solutions. The theoretical guarantees on the equivalence between the nuclear norm minimization and rank minimization problem were provided in [51, 36].

While the assumption of single clean structure is widely used, sometimes researchers deal with problems where the sparsity and the low-rank property are mixed together. Examples include the system identification problem with a sparse impulse response and small model order, and the Gaussian graphical model with latent and unobserved variables. Matrix decomposition models for these problems use the mix-structured regularization, which minimizes a weighted combination of the corresponding norm for the sparse structure and the nuclear norm for the low-rank matrix. The sparse structure in the mixture can be element-wise sparse [24], column-wise sparse [40, 58], or block-wise sparse [28].

1.2 Theoretical guarantees

There is a large number of theoretical results that guarantee the performance of various types of regularized M-estimators, such as the estimation of the error bound, the analysis of the prediction consistency, and the demonstration of the model consistency.

To begin with, we review some theoretical results for estimators with sparse constraints. The LASSO proposed in [53] is a popular technique for simultaneous estimation and model selection. Exact recovery for observations without noise has been discussed in [21, 39, 12]. The model selection consistency of the LASSO was investigated in [23, 67, 64, 42]. Moreover, the model selection consistency of Gaussian graphical models was studied in [41, 61]. Results for the model selection problem in the Gaussian concentration graph model, or the covariance selection problem were provided in [49, 8].

For the block-wise sparse model, the block-regularized estimator recovers the support of the model if the support is a union of groups and the covariates of different groups are not too correlated.

The grouped LASSO proposed in [61] is a popular block-regularized estimator with ℓ_1/ℓ_2-norm regularization. The main advantage of the group LASSO is that one can make a group of regression coefficients vanish simultaneously [34, 26], which is not possible for the LASSO. The group LASSO can be generalized to an infinite-dimensional setting [5]. Moreover, an extension of the group LASSO, called 'Blockwise Sparse Regression' (BSR) [33], works for general loss functions including generalized linear models. In addition, other variants of the group LASSO were studied and explored, such as the joint selection of covariates for multi-task learning [47].

In estimating low-rank matrices, theoretical guarantees for the equivalence between the nuclear norm minimization and rank minimization problem were provided in [51, 36]. The recovery of low-rank matrices was studied in [15] and the results were further improved in [13, 48, 50]. Necessary and sufficient conditions for rank consistency in noisy settings were provided by [4]. In practice, the nuclear norm minimization problem can be realized via semidefinite programming and there are different algorithms designed for solving this optimization problem, such as the interior point method [37], the gradient projection method, and the low-rank factorization technique. A low complexity algorithm which combines spectral techniques with manifold optimization was established in [31], and its performance guarantees were provided in [32]. Moreover, the strengths and weaknesses for different algorithms were studied in [51].

Furthermore, a line of research studied estimators based on the mixed structure. Most of the studies focus on the decomposition of the superposition of a low-rank component and a sparse component. The problem of decomposing the sum of a low-rank matrix and an entrywise sparse matrix was initially studied by [18], which demonstrated sufficient conditions for exact recovery with high probability based on the notion of rank-sparsity incoherence. Moreover, recovery given inaccurate samples was studied by [17, 24]. There are other studies which focus on the superposition of a column-wise sparse component and a low-rank component [40, 58], the superposition of a block-sparse component and an entry-wise sparse component [28], and so on. Furthermore, a general theory was developed in [1], which involves the above cases and yields non-asymptotic Frobenius error bounds for both deterministic and stochastic noise matrices.

1.3 Unified frameworks

In recent years, a line of on-going theoretical studies is focused on establishing a general framework for high-dimensional models, including some interesting scenarios as special cases. A unified framework was introduced by [45] to establish consistency and convergence rates for regularized M-estimators under high-dimensional scaling, under the assumption of the ”restricted strong convexity” on loss function and the decomposability for regularizers. This framework has also been used to prove several results on the estimation of low-rank matrices using the nuclear norm, as well as minimax-optimal rates for noisy matrix completion and noisy matrix decomposition. Moreover, a general framework for the high-dimensional analysis of "dirty" statistical models was established in [60], in which the model parameters are a superposition of structurally constrained parameters.

1.4 Contributions

In this paper, we establish a unified model for structure identification in high-dimensional problems. In our model's setting, two distinct structures are correlated by matrix coefficients. This framework can incorporate many models and applications. Special cases include principal component analysis problem and multiple regression problem. In this paper, we provide an innovative proof for the structural incoherence based on the estimation of the largest singular values of the product of two random matrices. In the deduction process, some results from probability theory and random matrix theory are used. Then, we establish error bounds that will hold with high probability. In addition, we illustrate our theoretical results via simulation for various parameters, that is, the sparse levels of the two unknown matrices which are to be recovered. Fortunately, the simulation results show that the property of structural incoherence of the coefficient matrices helps to reduce the error in recovering the correlated unknown sparse matrices. This good performance provides reliable evidence and verifications for theoretical analysis.

The remainder of this paper is organized as follows. In Chapter 2, we introduce the framework and then formulate it into a regularized least-squares problem. In Chapter 3, we derive some technical lemmas to demonstrate the main theorem concerning the error bound. In Chapter 4, we analyze the coefficient matrices of the specific model and demonstrate the structural incoherence with an innovative proof. In Chapter 5 we present simulation results and interpret their correspondence with the theoretical analysis. Finally we conclude the thesis via final remarks in Chapter 6.

Chapter 2

Problem Setup and Properties

2.1 Problem setup

We study the following general framework with mixed structures. We observe the response matrix $Y \in \mathbb{R}^{n\times q}$ and the covariate matrices $X_1 \in \mathbb{R}^{n\times p}, X_2 \in \mathbb{R}^{n\times n}$ such that

$$Y = X_1\Theta^* + X_2G^* + W. \tag{2.1}$$

Here $\Theta^* \in \mathbb{R}^{p\times q}$ and $G^* \in \mathbb{R}^{n\times q}$ represent some unknown linear relationships between the predictors X_1, X_2 and the response Y. And they enjoy some special low-dimensional structures, such as the element-wise sparse structure, the row-sparse structure or the low-rank structure. The matrix $W \in \mathbb{R}^{n\times q}$ is the noise matrix and its Frobenius norm is assumed to be small. The estimator is then written as

$$\begin{aligned}(\hat{\Theta}, \hat{G}) &\in \arg\min_{\Theta,G}\{\mathcal{L}(\Theta, G) + \lambda_1\mathcal{R}_1(\Theta) + \lambda_2\mathcal{R}_2(G)\} \\ &= \arg\min_{\Theta,G}\{\|Y - X_1\Theta - X_2G\|_F^2 + \lambda_1\mathcal{R}_1(\Theta) + \lambda_2\mathcal{R}_2(G)\}. \end{aligned}\tag{2.2}$$

The regularizers $\mathcal{R}_1, \mathcal{R}_2$ are determined based on the assumed structures in the corresponding specific setting.

Generally speaking, based on the properties of X_1 and X_2, we can split the general model into 4 cases.

First Case: Both X_1 and X_2 are deterministic. For example, the PCA model [1, 58], $Y = \Theta^* + G^* + W$, where $W \in \mathbb{R}^{n\times n}$ is a Wishart distributed matrix, $\Theta^* \in \mathbb{R}^{n\times n}$ is low-rank, and $G^* \in \mathbb{R}^{n\times n}$ is entry-wise sparse. Here in this example $X_1 = X_2 = I_{n\times n}$. The estimator is

$$(\hat{\Theta}, \hat{G}) \in \arg\min_{\Theta,G}\{\|Y - \Theta - G\|_F^2 + \lambda_1\|\Theta\|_* + \lambda_2\|G\|_1\}. \tag{2.3}$$

In this paper, $\|.\|_*$ denotes the nuclear norm while $\|.\|_1$ stands for the entry-wise ℓ_1-norm.

Second Case: For X_1 and X_2, one of them is deterministic while another one is a random matrix. For example, the robust multi-task model with corrupted gross errors [59], $Y = X_1\Theta^* + G^* + W$, where G^* represents the gross error and Θ^* is assumed to be row-sparse. In this example $X_2 = I_{n\times n}$ and X_1 is a random matrix with sub-Gaussian rows. The estimator is in the following form:

$$(\hat{\Theta}, \hat{G}) \in \arg\min_{\Theta,G}\{\|Y - X_1\Theta - G\|_F^2 + \lambda_1\|\Theta\|_{1,2} + \lambda_2\|G\|_1\}. \tag{2.4}$$

Note that $\|.\|_{1,2}$ stands for the ℓ_1/ℓ_2-norm that is the sum of ℓ_2-norm of rows of the matrix.

Third Case: Both X_1 and X_2 are random matrices and they are correlated. For instance, the multi-linear regression model with mixed structure: $Y = X(\Theta^* + G^*) + W$. In this case, $X_1 = X_2 = X$. The two random matrices X_1 and X_2 are fully correlated.

Fourth Case: X_1 and X_2 are independent random matrices.

In the following chapters, we mainly discuss the four properties and the error bound for the last case where the coefficient matrices X_1 and X_2 are correlated and independent random matrices. In fact, it is easier to study the first three cases [60, 59, 1, 58]. In Section 4.3 we will derive similar results for the first three cases by deriving corollaries based on the main theorem.

2.2 Assumptions and notations

Set $Z := X_1\Theta + X_2G$, then $Z^* = X_1\Theta^* + X_2G^*$, and the small deviation Δ_Z equals to $X_1\Delta_\Theta + X_2\Delta_G$. Then we can write the loss function as a function of Z: $\mathcal{L}(Z)$. Define the optimal errors $\hat{\Delta}_\Theta = \hat{\Theta} - \Theta^*$ and $\hat{\Delta}_G = \hat{G} - G^*$.

Let's state some natural assumptions on the regularization functions $\mathcal{R}_\alpha$ ($\alpha = 1, 2$) and the loss function $\mathcal{L}$ for model (2.1).

(A1) The loss function $\mathcal{L}$ is convex and differentiable.

(A2) The regularizers $\mathcal{R}_\alpha$ are norms and are decomposable with respect to the subspace pairs $(\mathcal{M}_\alpha, \overline{\mathcal{M}}_\alpha^\perp)$, where $\mathcal{M}_\alpha \in \overline{\mathcal{M}}_\alpha$.

Remark. *Decomposability means $\mathcal{R}_\alpha(u+v) = \mathcal{R}_\alpha(u) + \mathcal{R}_\alpha(v)$ for all $u \in \mathcal{M}_\alpha, v \in \overline{\mathcal{M}_\alpha}^\perp$. $\mathcal{M}_\alpha$ is the corresponding low-dimensional subspace. The property of decomposition of a regularization function captures the suitability of a regularizer to a particular structure.*

Our next requirement is the "restricted strong convexity" [45].

(A3) [Restricted Strong Convexity]

$$\delta\mathcal{L}(\Delta_\Theta; \Theta^*, G^*) \geq \mathcal{K}_L\|\Delta_\Theta\|_F^2 - \mathcal{G}_1\mathcal{R}_1^2(\Delta_\Theta), \tag{2.5}$$

$$\delta\mathcal{L}(\Delta_G; \Theta^*, G^*) \geq \mathcal{K}_L\|\Delta_G\|_F^2 - \mathcal{G}_2\mathcal{R}_2^2(\Delta_G). \tag{2.6}$$

Note that the assumptions (A1)-(A3) are usually imposed on the model even when there is only one clean structure. Our next assumption is on the interaction between the different structured items [60] in mix-structure models.

(A4) [Structural Incoherence]

$$\begin{aligned}&|\mathcal{L}(Z^* + X_1\Delta_\Theta + X_2\Delta_G) + \mathcal{L}(Z^*) - \mathcal{L}(Z^* + X_1\Delta_\Theta) - \mathcal{L}(Z^* + X_2\Delta_G)|\\ \leq &\frac{\mathcal{K}_L}{2}(\|\Delta_\Theta\|_F^2 + \|\Delta_G\|_F^2) + \sum_{\alpha=1,2}\mathcal{H}_\alpha\mathcal{R}_\alpha^2(\Delta_\alpha). \end{aligned}\tag{2.7}$$

Chapter 3

Estimation of the Error Bound

In this chapter, we estimate the error bound for the optimization problem (2.2), based on the four natural assumptions (A1)-(A4). Firstly we present the main result (Theorem 3.1) in Section 3.1. We will provide the proof in Section 3.3.

3.1 Main theorem

Note that the theorem involves the concept of subspace compatibility constant $\Psi(\mathcal{M}, \|.\|) := sup_{u\in\mathcal{M}\{0\}}\frac{\mathcal{R}}{\|u\|}$, defined in [45]. This notion captures the relationship between the regularization function $\mathcal{R}(.)$ and the error norm $\|.\|$ over vectors in the subspace, and it will be widely used in the following results.

Theorem 3.1. *Suppose that (A1)-(A4) are satisfied. Define optimal error* $\hat{\Delta}_G = \hat{G} - G^*$*,* $\hat{\Delta}_\Theta = \hat{\Theta} - \Theta^*$*. Then we have:*

$$\|\hat{\Theta} - \Theta^*\|_F + \|\hat{G} - G^*\|_F \leq \frac{3\Phi + 2\sqrt{\mathcal{K}D}}{\overline{\mathcal{K}}}, \tag{3.1}$$

where

$$\begin{aligned}\Phi =& \max\{\lambda_1\Psi_1(\overline{\mathcal{M}_1}), \lambda_2\Psi_2(\overline{\mathcal{M}_2})\},\\ 2D =&\overline{\tau}(Z^*) + 2\lambda_1\mathcal{R}_1(\Pi_{\overline{\mathcal{M}}_1^\perp}\Theta^*) + 2\lambda_2\mathcal{R}_2(\Pi_{\overline{\mathcal{M}}_2^\perp}G^*),\\ \overline{\mathcal{K}} =& \frac{\mathcal{K}_L}{2} - 64\overline{\mathcal{G}}^2\Phi^2,\\ \overline{\tau}(Z^*) =& 64\overline{\mathcal{G}}^2(\lambda_1^2\mathcal{R}_1(\Pi_{\overline{\mathcal{M}}_1^\perp}(\Theta^*))^2 + \lambda_2^2\mathcal{R}_2(\Pi_{\overline{\mathcal{M}}_2^\perp}(G^*))^2),\\ \overline{\mathcal{G}} =& max_{\alpha=1,2}\frac{1}{\lambda_\alpha}\sqrt{\mathcal{G}_\alpha + \mathcal{H}_\alpha}.\end{aligned}$$

3.2 Preliminaries

In this section, we provide some technical lemmas to build up the theoretical base for the proof of Theorem 3.1. For the convenience of proofs, we list some notions and define some functions. In the following context, we will use $\Pi_{\mathcal{M}}(A)$ to denote the projection of matrix A onto the subspace $\mathcal{M}$. Define the set

$$\begin{aligned}\mathbb{C} :=& \{(\Delta_\Theta, \Delta_G) : \lambda_1\mathcal{R}_1(\Pi_{\overline{\mathcal{M}}_1^\perp}(\Delta_\Theta)) + \lambda_2\mathcal{R}_2(\Pi_{\overline{\mathcal{M}}_2^\perp}(\Delta_G)) \\ &\leq \lambda_1[3\mathcal{R}_1(\Pi_{\overline{\mathcal{M}}_1}(\Delta_\Theta)) + 4\mathcal{R}_1(\Pi_{\overline{\mathcal{M}}_1}(\Theta^*))] + \lambda_2[3\mathcal{R}_2(\Pi_{\overline{\mathcal{M}}_2}(\Delta_G)) + 4\mathcal{R}_2(\Pi_{\overline{\mathcal{M}}_2}(G^*))]\}.\end{aligned} \tag{3.2}$$

Define

$$\delta\mathcal{L}(\Delta_Z) := \mathcal{L}(Z^* + \Delta Z) - \mathcal{L}(Z^*) - \langle\nabla_Z\mathcal{L}(Z^*), X_1\Delta_\Theta\rangle - \langle\nabla_Z\mathcal{L}(Z^*), X_2\Delta_G\rangle, \tag{3.3}$$

$$\begin{aligned}F(\Delta_\Theta, \Delta_G) :=& \mathcal{L}(Z^* + \Delta_Z) - \mathcal{L}(Z^*)\\ &+ \lambda_1[\mathcal{R}_1(\Theta^* + \Delta_\Theta) - \mathcal{R}_1(\Theta^*)] + \lambda_2[\mathcal{R}_2(G^* + \Delta_G) - \mathcal{R}_2(G^*)].\end{aligned} \tag{3.4}$$

We can see that $F(\Delta_\Theta, \Delta_G)$ is the difference of the objective function values at $Z^* + \Delta_Z$ and Z^*. We can rewrite it as

$$\begin{aligned}F(\Delta_\Theta, \Delta_G) =& \delta L(\Delta_Z) + \langle\nabla_Z\mathcal{L}(Z^*), X_1\Delta_\Theta\rangle + \langle\nabla_Z\mathcal{L}(Z^*), X_2\Delta_G\rangle\\ &+ \lambda_1[\mathcal{R}_1(\Theta^* + \Delta_\Theta) - \mathcal{R}_1(\Theta^*)] + \lambda_2[\mathcal{R}_2(G^* + \Delta_G) - \mathcal{R}_2(G^*)].\end{aligned} \tag{3.5}$$

Remark. *Define the optimal deviations* $\hat{\Delta}_\Theta = \hat{\Theta} - \Theta^*, \hat{\Delta}_G = \hat{G} - G^*$*, and* $\hat{\Delta}_Z = X_1\hat{\Delta}_\Theta + X_2\hat{\Delta}_G = \hat{Z} - Z^*$*. Then*

$$F(\hat{\Delta}_\Theta, \hat{\Delta}_G) = Objectivefunction(\hat{Z}) - Objectivefunction(Z^*) \leq 0. \tag{3.6}$$

In particular, $F(0,0) = 0$*,*

Now let's start with the following lemma which will be used in the following results.

Lemma 3.2 (Deviation inequalities)**.** *For any decomposable regularizer, we have*

$$\mathcal{R}(\theta^* + \Delta) - \mathcal{R}(\theta^*) \geq \mathcal{R}(\Pi_{\overline{\mathcal{M}}^\perp}(\Delta)) - \mathcal{R}(\Pi_{\overline{\mathcal{M}}}(\Delta)) - 2\mathcal{R}(\Pi_{\mathcal{M}^\perp}(\theta^*)), \tag{3.7}$$

where θ^* *and* Δ *are p-dimensional vectors.*
Moreover, as long as $\lambda \geq 2\mathcal{R}^*(\nabla\mathcal{L}(\theta^*))$ *and* $\mathcal{L}$ *is convex, we have*

$$\mathcal{L}(\theta^* + \Delta) - \mathcal{L}(\theta^*) \geq -\frac{\lambda}{2}[\mathcal{R}(\Pi_{\overline{\mathcal{M}}}(\Delta)) + \mathcal{R}(\Pi_{\overline{\mathcal{M}}^\perp}(\Delta))]. \tag{3.8}$$

The detailed proof can be found in [60] appendix

Next, we present two technical results concerning the estimation of the optimal error under the assumptions (A1) and (A2).

Lemma 3.3. *Suppose that (A1) and (A2) are satisfied,* $\lambda_1 \geq 2\mathcal{R}_1^*(X_1^T\nabla_Z\mathcal{L}(Z^*))$*, and* $\lambda_2 \geq 2\mathcal{R}_2^*(X_2^T\nabla_Z\mathcal{L}(Z^*))$*. Then the optimal error* $\hat{\Delta}$ *lies in the set*

$$\begin{aligned}\mathbb{C} :=& \{(\Delta_\Theta, \Delta_G) : \lambda_1\mathcal{R}_1(\Pi_{\overline{\mathcal{M}}_1^\perp}(\Delta_\Theta)) + \lambda_2\mathcal{R}_2(\Pi_{\overline{\mathcal{M}}_2^\perp}(\Delta_G)) \\ &\leq \lambda_1[3\mathcal{R}_1(\Pi_{\overline{\mathcal{M}}_1}(\Delta_\Theta)) + 4\mathcal{R}_1(\Pi_{\overline{\mathcal{M}}_1}(\Theta^*))] + \lambda_2[3\mathcal{R}_2(\Pi_{\overline{\mathcal{M}}_2}(\Delta_G)) + 4\mathcal{R}_2(\Pi_{\overline{\mathcal{M}}_2}(G^*))]\}.\end{aligned}$$

Proof. In this proof, we make use of the function $F(\Delta_\Theta, \Delta_G)$ to achieve the conclusion.
Recall the definition of $F(\Delta_\Theta, \Delta_G)$ (3.3), which is

$$\begin{aligned}F(\Delta_\Theta, \Delta_G) :=& \mathcal{L}(Z^* + \Delta_Z) - \mathcal{L}(Z^*) \\ &+ \lambda_1[\mathcal{R}_1(\Theta^* + \Delta_\Theta) - \mathcal{R}_1(\Theta^*)] + \lambda_2[\mathcal{R}_2(G^* + \Delta_G) - \mathcal{R}_2(G^*)].\end{aligned}$$

We know

$$F(\hat{\Delta}_\Theta, \hat{\Delta}_G) := Objectivefn(\hat{Z}) - Objectivefn(Z^*) \leq 0.$$

Applying Lemma (3.2) for decomposable regularizers $\mathcal{R}_1$ and $\mathcal{R}_2$, and get:

$$\begin{aligned} &\mathcal{R}_1(\Theta^* + \Delta_\Theta) - \mathcal{R}_1(\Theta^*) \\ \geq &\mathcal{R}_1(\Pi_{\overline{\mathcal{M}_1}^\perp}(\Delta_\Theta)) - \mathcal{R}_1(\Pi_{\overline{\mathcal{M}_1}}(\Delta_\Theta)) - 2\mathcal{R}(\Pi_{\mathcal{M}_1^\perp}(\Theta^*)), \end{aligned} \tag{3.9}$$

$$\begin{aligned} &\mathcal{R}_2(G^* + \Delta_G) - \mathcal{R}_2(G^*) \\ \geq &\mathcal{R}_2(\Pi_{\overline{\mathcal{M}_2}^\perp}(\Delta_G)) - \mathcal{R}_2(\Pi_{\overline{\mathcal{M}_2}}(\Delta_G)) - 2\mathcal{R}_2(\Pi_{\mathcal{M}_2^\perp}(G^*)). \end{aligned} \tag{3.10}$$

Moreover, as long as $\lambda \geq 2\mathcal{R}^*(\nabla\mathcal{L}(\theta^*))$ and $\mathcal{L}$ is convex, we have

$$\begin{aligned} &\mathcal{L}(Z^* + \Delta Z) - \mathcal{L}(Z^*) \\ =&\mathcal{L}(X_1\Theta^* + X_2G^* + X_1\Delta\Theta + X_2\Delta G) - \mathcal{L}(X_1\Theta^* + X_2G^*) \\ \geq&\langle \nabla_\Theta\mathcal{L}(Z^*), \Delta_\Theta\rangle + \langle \nabla_G\mathcal{L}(Z^*), \Delta_G\rangle \\ \overset{(i)}{\geq}& - \mathcal{R}_1^*(\nabla_\Theta\mathcal{L}(Z^*))\mathcal{R}_1(\Delta_\Theta) - \mathcal{R}_2^*(\nabla_G\mathcal{L}(Z^*))\mathcal{R}_2(\Delta_G) \\ \overset{(ii)}{\geq}& - \frac{\lambda_1}{2}[\mathcal{R}_1(\Pi_{\overline{\mathcal{M}_1}}(\Delta_\Theta)) + \mathcal{R}_1(\Pi_{\overline{\mathcal{M}_1}^\perp}(\Delta_\Theta))] \\ &- \frac{\lambda_2}{2}[\mathcal{R}_2(\Pi_{\overline{\mathcal{M}_2}}(\Delta_G)) + \mathcal{R}_2(\Pi_{\overline{\mathcal{M}_2}^\perp}(\Delta_G))]. \end{aligned} \tag{3.11}$$

where the inequality (i) is from the generalized Cauchy-Schwarz inequality, and the inequality (ii) is based on the decomposability of the regularizers and the assumptions on λ in the statement of this lemma.

Combining (3.9), (3.10) and (3.11), we obtain

$$
\begin{aligned}
0 \geq F(\hat{\Delta}_\Theta, \hat{\Delta}_G) =& \mathcal{L}(Z^* + \hat{\Delta}_Z) - \mathcal{L}(Z^*) \\
& + \lambda_1[\mathcal{R}_1(\Theta^* + \hat{\Delta}_\Theta) - \mathcal{R}_1(\Theta^*)] + \lambda_2[\mathcal{R}_2(G^* + \hat{\Delta}_G) - \mathcal{R}_2(G^*)] \\
\geq & -\frac{\lambda_1}{2}[\mathcal{R}_1(\Pi_{\overline{\mathcal{M}_1}}(\Delta_\Theta)) + \mathcal{R}_1(\Pi_{\overline{\mathcal{M}_1}^\perp}(\Delta_\Theta))] \\
& + \lambda_1[\mathcal{R}_1(\Pi_{\overline{\mathcal{M}_1}^\perp}(\Delta_\Theta)) - \mathcal{R}_1(\Pi_{\overline{\mathcal{M}_1}}(\Delta_\Theta)) - 2\mathcal{R}(\Pi_{\mathcal{M}_1^\perp}(\Theta^*))] \\
& - \frac{\lambda_2}{2}[\mathcal{R}_2(\Pi_{\overline{\mathcal{M}_2}}(\Delta_G)) + \mathcal{R}_2(\Pi_{\overline{\mathcal{M}_2}^\perp}(\Delta_G))]. \\
& + \lambda_2[\mathcal{R}_2(\Pi_{\overline{\mathcal{M}_2}^\perp}(\Delta_G)) - \mathcal{R}_2(\Pi_{\overline{\mathcal{M}_2}}(\Delta_G)) - 2\mathcal{R}_2(\Pi_{\mathcal{M}_2^\perp}(G^*))] \\
= & -\frac{3\lambda_1}{2}\mathcal{R}_1(\Pi_{\overline{\mathcal{M}_1}}(\Delta_\Theta)) + \frac{\lambda_1}{2}\mathcal{R}_1(\Pi_{\overline{\mathcal{M}_1}^\perp}(\Delta_\Theta)) - 2\lambda_1\mathcal{R}(\Pi_{\mathcal{M}_1^\perp}(\Theta^*)) \\
& - \frac{3\lambda_2}{2}\mathcal{R}_2(\Pi_{\overline{\mathcal{M}_2}}(\Delta_\Theta)) + \frac{\lambda_2}{2}\mathcal{R}_2(\Pi_{\overline{\mathcal{M}_2}^\perp}(\Delta_\Theta)) - 2\lambda_2\mathcal{R}(\Pi_{\mathcal{M}_2^\perp}(\Theta^*)).
\end{aligned}
$$

A simple reformulation completes the proof. □

Lemma 3.4. *Suppose that (A1) and (A2) are satisfied. $\mathbb{C}$ defined as in equation (3.2). If $F(\Delta_\Theta, \Delta_G) > 0$ for all possible vectors $(\Delta_\Theta, \Delta_G) \in K(\delta) := \mathbb{C} \cap \{\|\Delta_\Theta\| + \|\Delta_G\| = \delta\}$, then the optimal error satisfies*

$$\|\hat{\Delta}_\Theta\|_F + \|\hat{\Delta}_G\|_F \leq \delta, \tag{3.12}$$

where $\hat{\Delta}_\Theta = \hat{\Theta} - \Theta^*$, and $\hat{\Delta}_G = \hat{G} - G^*$.

Actually, a similar result was provided in [60]. We present the proof here as a reference.

Proof. Let's first show some special property of the set $\mathbb{C}$, based on which we will then derive the guarantees for error bound.

Let $(\Delta_\Theta, \Delta_G)$ be an arbitrary error vector in the set $\mathbb{C}$. Then, for any $t \in (0, 1)$, we have

$$
\begin{aligned}
&\lambda_1\mathcal{R}_1(\Pi_{\overline{\mathcal{M}_1}^\perp}(t\Delta_\Theta)) + \lambda_2\mathcal{R}_2(\Pi_{\overline{\mathcal{M}_2}^\perp}(t\Delta_G)\|)\\
\overset{(i)}{=}&\lambda_1\mathcal{R}_1(t\Pi_{\overline{\mathcal{M}_1}^\perp}(\Delta_\Theta)) + \lambda_2\mathcal{R}_2(t\Pi_{\overline{\mathcal{M}_2}^\perp}(\Delta_G))\\
\overset{(ii)}{=}&t\lambda_1\mathcal{R}_1(\Pi_{\overline{\mathcal{M}_1}^\perp}(\Delta_\Theta)) + t\lambda_2\mathcal{R}_2(\Pi_{\overline{\mathcal{M}_2}^\perp}(\Delta_G))\\
\overset{(iii)}{\le}&t\lambda_1[3\mathcal{R}_1(\Pi_{\overline{\mathcal{M}_1}}(\Delta_\Theta)) + 4\mathcal{R}_1(\Pi_{\mathcal{M}_1^\perp}(\Theta^*))] + t\lambda_2[3\mathcal{R}_2(\Pi_{\overline{\mathcal{M}_2}}(\Delta_G)) + 4\mathcal{R}_2(\Pi_{\mathcal{M}_2^\perp}(G^*))]\\
\overset{(iv)}{=}&\lambda_1[3\mathcal{R}_1(\Pi_{\overline{\mathcal{M}_1}}(t\Delta_\Theta)) + 4t\mathcal{R}_1(\Pi_{\mathcal{M}_1^\perp}(\Theta^*))] + \lambda_2[3\mathcal{R}_2(\Pi_{\overline{\mathcal{M}_2}}(t\Delta_G)) + 4t\mathcal{R}_2(\Pi_{\mathcal{M}_2^\perp}(G^*))]\\
\overset{(v)}{\le}&\lambda_1[3\mathcal{R}_1(\Pi_{\overline{\mathcal{M}_1}}(t\Delta_\Theta)) + 4\mathcal{R}_1(\Pi_{\mathcal{M}_1^\perp}(\Theta^*))] + \lambda_2[3\mathcal{R}_2(\Pi_{\overline{\mathcal{M}_2}}(t\Delta_G)) + 4\mathcal{R}_2(\Pi_{\mathcal{M}_2^\perp}(G^*))]
\end{aligned}
$$

where step (i) uses the fact that

$$
\begin{aligned}
\Pi_{\overline{\mathcal{M}_1}^\perp}(t\Delta_\Theta) =&\arg\min{}_{\gamma\in\overline{\mathcal{M}}^\perp}\|t\Delta_\Theta - \gamma\|\\
=&t\arg\min{}_{\gamma\in\overline{\mathcal{M}}^\perp}\|\Delta_\Theta - \frac{\gamma}{t}\|\\
=&t\Pi_{\overline{\mathcal{M}_1}^\perp}(\Delta_\Theta)\\
\Pi_{\overline{\mathcal{M}_2}^\perp}(t\Delta_G) =&\arg\min{}_{\gamma\in\overline{\mathcal{M}}^\perp}\|t\Delta_G - \gamma\|\\
=&t\arg\min{}_{\gamma\in\overline{\mathcal{M}}^\perp}\|\Delta_G - \frac{\gamma}{t}\|\\
=&\ t\Pi_{\overline{\mathcal{M}_2}^\perp}(\Delta_G),
\end{aligned}
$$

and step (ii) uses the positive homogeneity of norms, and step (iii) holds since $(\Delta_\Theta, \Delta_G) \in \mathbb{C}$.

Moreover, step (iv) holds similarly as in equalities (i) and (ii), and finally step (v) trivially holds for any $t \le 1$.

Therefore, if $(\Delta_\Theta, \Delta_G) \in \mathbb{C}$, then the line segment $\{(t\Delta_\Theta, t\Delta_G), t \in (0,1)\}$ between $(\Delta_\Theta, \Delta_G)$ and (0,0) also lies in $\mathbb{C}$.

Suppose $\|\hat{\Delta}_\Theta\|_F + \|\hat{\Delta}_G\|_F > \delta$. Since $\|t\hat{\Delta}_\Theta\|_F + \|t\hat{\Delta}_G\|_F = t\|\hat{\Delta}_\Theta\|_F + t\|\hat{\Delta}_G\|_F$, there exists some constant $t^* \in (0,1)$ s.t $(t^*\hat{\Delta}_\Theta, t^*\hat{\Delta}_G) \in K(\delta)$. At the same time, by the convexity of $\mathcal{L}$ and the regularizers,

$$F(t^*\hat{\Delta}_\Theta, t^*\hat{\Delta}_G) \le t^*F(\hat{\Delta}_\Theta, \hat{\Delta}_G) + (1-t^*)F(0,0) \le 0.$$

Therefore, $(t^*\hat{\Delta}_\Theta, t^*\hat{\Delta}_G)$ is in $K(\delta)$ such that $F(t^*\hat{\Delta}_\Theta, t^*\hat{\Delta}_G) \le 0$ by construction. Hence the statement follows. □

In the following context, we show that (A3) and (A4) indicate the global restricted strong convexity which bounds δL with a nice estimate. Actually, this lemma and its proof have roots in [60].

Lemma 3.5. *Suppose that (A3) and (A4) are satisfied. Then the global RSC (Restricted Strong Convexity) holds:*

$$\begin{aligned}\delta\mathcal{L} :=& \mathcal{L}(Z^* + \Delta Z) - \mathcal{L}(Z^*) - \langle \nabla_\Theta \mathcal{L}(Z^*), \Delta_\Theta \rangle - \langle \nabla_G \mathcal{L}(Z^*), \Delta_G \rangle \\ \geq& \overline{\mathcal{K}}(\|\Delta_\Theta\|_F^2 + \|\Delta_G\|_F^2) - \overline{\tau}(Z^*),\end{aligned} \tag{3.13}$$

where

$$\begin{aligned}\overline{\mathcal{K}} =& \frac{\mathcal{K}_L}{2} - 64\overline{\mathcal{G}}^2\Phi^2, \\ \overline{\tau}(Z^*) =& 64\overline{g}^2(\lambda_1^2\mathcal{R}_1^2(\Pi_{\overline{\mathcal{M}}_1^\perp}(\Theta^*)) + \lambda_2^2\mathcal{R}_2^2(\Pi_{\overline{\mathcal{M}}_2^\perp}(G^*))), \\ \overline{\mathcal{G}} =& max_{\alpha=1,2}\frac{1}{\lambda_\alpha}\sqrt{\mathcal{G}_\alpha + \mathcal{H}_\alpha}.\end{aligned}$$

Proof. Observing the composition of $\delta\mathcal{L}(\Delta_Z)$, we split it into three parts and bound every term separately.

$$\begin{aligned}\delta\mathcal{L}(\Delta_Z) \ :=& \mathcal{L}(Z^* + \Delta Z) - \mathcal{L}(Z^*) - \langle \nabla_Z \mathcal{L}(Z^*), X_1\Delta_\Theta \rangle - \langle \nabla_Z \mathcal{L}(Z^*), X_2\Delta_G \rangle, \\ =& \mathcal{L}(Z^* + X_1\Delta_\Theta + X_2\Delta_G) + \mathcal{L}(Z^*) - \mathcal{L}(Z^* + X_1\Delta_\Theta) - \mathcal{L}(Z^* + X_2\Delta_G) \\ &+ [\mathcal{L}(Z^* + X_1\Delta_\Theta) - \mathcal{L}(Z^*) - \langle \nabla_Z \mathcal{L}(Z^*), X_1\Delta_\Theta \rangle] \\ &+ [\mathcal{L}(Z^* + X_2\Delta_G) - \mathcal{L}(Z^*) - \langle \nabla_Z \mathcal{L}(Z^*), X_2\Delta_G \rangle] \\ :=& W_1 + W_2 + W_3,\end{aligned}$$

where

$$\begin{aligned}W_1 =& \mathcal{L}(Z^* + X_1\Delta_\Theta + X_2\Delta_G) + \mathcal{L}(Z^*) - \mathcal{L}(Z^* + X_1\Delta_\Theta) - \mathcal{L}(Z^* + X_2\Delta_G), \\ W_2 =& [\mathcal{L}(Z^* + X_1\Delta_\Theta) - \mathcal{L}(Z^*) - \langle \nabla_Z \mathcal{L}(Z^*), X_1\Delta_\Theta \rangle], \\ W_3 =& [\mathcal{L}(Z^* + X_2\Delta_G) - \mathcal{L}(Z^*) - \langle \nabla_Z \mathcal{L}(Z^*), X_2\Delta_G \rangle].\end{aligned}$$

By the inequalities (2.5), (2.6) and (2.7), we get

$$\begin{aligned}
W_1 \geq & -\frac{\mathcal{K}_L}{2}(\|\Delta_\Theta\|_F^2 + \|\Delta_G\|_F^2) - \sum_{\alpha=1,2} \mathcal{H}_\alpha \mathcal{R}_\alpha^2(\Delta_\alpha), \\
W_2 \geq & \mathcal{K}_L \|\Delta_\Theta\|_F^2 - \mathcal{G}_1 \mathcal{R}_1(\Delta_\Theta)^2, \\
W_3 \geq & \mathcal{K}_L \|\Delta_G\|_F^2 - \mathcal{G}_2 \mathcal{R}_2(\Delta_G)^2.
\end{aligned}$$

Then,

$$\begin{aligned}
\delta\mathcal{L}(\Delta_Z) = & W_1 + W_2 + W_3 \\
\geq & \mathcal{K}_L \|\Delta_\Theta\|_F^2 - \mathcal{G}_1 \mathcal{R}_1^2(\Delta_\Theta) + \mathcal{K}_L \|\Delta_G\|_F^2 - \mathcal{G}_2 \mathcal{R}_2^2(\Delta_G) \\
& - \frac{\mathcal{K}_L}{2}(\|\Delta_\Theta\|_F^2 + \|\Delta_G\|_F^2) - \mathcal{H}_1 \mathcal{R}_1^2(\Delta_\Theta) - \mathcal{H}_2 \mathcal{R}_2^2(\Delta_G) \\
= & \frac{\mathcal{K}_L}{2}(\|\Delta_\Theta\|_F^2 + |\Delta_G\|_F^2) - (\mathcal{G}_1 + \mathcal{H}_1)\mathcal{R}_1^2(\Delta_\Theta) - (\mathcal{G}_2 + \mathcal{H}_2)\mathcal{R}_2^2(\Delta_G) \\
:= & \frac{\mathcal{K}_L}{2}(\|\Delta_\Theta\|_F^2 + |\Delta_G\|_F^2) - U,
\end{aligned}$$

where

$$U = (\mathcal{G}_1 + \mathcal{H}_1)\mathcal{R}_1^2(\Delta_\Theta) + (\mathcal{G}_2 + \mathcal{H}_2)\mathcal{R}_2^2(\Delta_G).$$

And we also have

$$\begin{aligned}
U = & (\mathcal{G}_1 + \mathcal{H}_1)\mathcal{R}_1^2(\Delta_\Theta) + (\mathcal{G}_2 + \mathcal{H}_2)\mathcal{R}_2^2(\Delta_G) \\
\leq & (\sqrt{\mathcal{G}_1 + \mathcal{H}_1}\ \mathcal{R}_1(\Delta_\Theta) + \sqrt{\mathcal{G}_2 + \mathcal{H}_2}\mathcal{R}_2(\Delta_G)^2 \\
\leq & [\overline{\mathcal{G}}(\lambda_1 \mathcal{R}_1(\Delta_\Theta) + \lambda_2 \mathcal{R}_2(\Delta_G))]^2,
\end{aligned}$$

where in the second inequality we use $\langle x, y\rangle \leq \|x\|_\infty \|y\|_1$ and $\overline{\mathcal{G}} := max_{\alpha=1,2} \frac{1}{\lambda_\alpha}\sqrt{\mathcal{G}_\alpha + \mathcal{H}_\alpha}$. By Lemma 3.3, for any $(\Delta_\Theta, \Delta_G) \in \mathbb{C}$,

$$\begin{aligned}
& \lambda_1 \mathcal{R}_1(\Delta_\Theta) + \lambda_2 \mathcal{R}_2(\Delta_G) \\
\leq & \lambda_1(\mathcal{R}_1(\Pi_{\overline{\mathcal{M}}_1}(\Delta_\Theta)) + \mathcal{R}_1(\Pi_{\overline{\mathcal{M}}_1^\perp}(\Delta_\Theta))) + \lambda_2(\mathcal{R}_2(\Pi_{\overline{\mathcal{M}}_2}(\Delta_G)) + \mathcal{R}_2(\Pi_{\overline{\mathcal{M}}_2^\perp}(\Delta_G))) \\
\leq & \lambda_1(4\mathcal{R}_1 \Pi_{\overline{\mathcal{M}}_1}(\Delta_\Theta)) + 4\mathcal{R}_1(\Pi_{\overline{\mathcal{M}}_1^\perp}(\Theta^*))) + \lambda_2(4\mathcal{R}_2(\Pi_{\overline{\mathcal{M}}_2}(\Delta_G)) + 4\mathcal{R}_2(\Pi_{\overline{\mathcal{M}}_2^\perp}(G^*))).
\end{aligned}$$

Therefore,

$$\begin{aligned}
U \leq & \overline{\mathcal{G}}^2 [\lambda_1 \mathcal{R}_1(\Theta) + \lambda_2 \mathcal{R}_2(\Delta_G)]^2 \\
\leq & \overline{\mathcal{G}}^2 16 \times 4 \times [\lambda_1^2 \mathcal{R}_1(\Pi_{\overline{\mathcal{M}}_1}(\Delta_\Theta))^2 + \lambda_1^2 \mathcal{R}_1(\Pi_{\overline{\mathcal{M}}_1^\perp}(\Theta^*))^2 + \lambda_2^2 \mathcal{R}_2(\Pi_{\overline{\mathcal{M}}_2}(\Delta_G))^2 + \lambda_2^2 \mathcal{R}_2(\Pi_{\overline{\mathcal{M}}_2^\perp}(G^*))^2] \\
\leq & 64 \overline{\mathcal{G}}^2 [\lambda_1^2 \Psi_1^2 \|\Delta_\Theta\|_F^2 + \lambda_1^2 \mathcal{R}_1(\Pi_{\overline{\mathcal{M}}_1^\perp}(\Theta^*))^2 + \lambda_2^2 \Psi_2^2 |\Delta_G\|_F^2 + \lambda_2^2 \mathcal{R}_2(\Pi_{\overline{\mathcal{M}}_2^\perp}(G^*))^2].
\end{aligned}$$

Thus we get the lower-bound for $\delta\mathcal{L}$:

$$\begin{aligned}
& \delta\mathcal{L}(\Delta_Z; Z^*) \\
\geq & \frac{\mathcal{K}_L}{2}(\|\Delta_\Theta\|_F^2 + |\Delta_G\|_F^2) - U \\
\geq & \frac{\mathcal{K}_L}{2}(\|\Delta_\Theta\|_F^2 + |\Delta_G\|_F^2) - 64\overline{\mathcal{G}}^2(\lambda_1^2 \Psi_1^2 \|\Delta_\Theta\|_F^2 + \lambda_1^2 \mathcal{R}_1(\Pi_{\overline{\mathcal{M}}_1^\perp}(\Theta^*))^2 \\
& + \lambda_2^2 \Psi_2^2 \|\Delta_G\|_F^2 + \lambda_2^2 \mathcal{R}_2(\Pi_{\overline{\mathcal{M}}_2^\perp}(G^*))^2) \\
\geq & (\frac{\mathcal{K}_L}{2} - 64\overline{\mathcal{G}}^2 \Phi^2)(\|\Delta_\Theta\|_F^2 + \|\Delta_G\|_F^2) - 64\overline{\mathcal{G}}^2(\lambda_1^2 \mathcal{R}_1(\Pi_{\overline{\mathcal{M}}_1^\perp}(\Theta^*))^2 + \lambda_2^2 \mathcal{R}_2(\Pi_{\overline{\mathcal{M}}_2^\perp}(G^*))^2),
\end{aligned}$$

which concludes the proof. □

Besides the above results, we still need to estimate the regularization part of the objective function. In fact, we will use the following lemma to attain the minimum in the estimation of $F(\Delta_\Theta, \Delta_G)$ in Section 3.3.

Lemma 3.6. *Consider the 2-variate quadratic function: $F(x_1, x_2) = ax_1^2 + bx_1 + c + ax_2^2 + bx_2 + c$ for some constants a, b,c. In addition, we suppose $a > 0, x_1 \geq 0, x_2 \geq 0x_1 + x_2 = \delta > 0$. Then $F(x_1, x_2)$ attains its minimum value at $x_1 = x_2 = \delta/2$.*

3.3 Demonstration of the main theorem

In this section, we employ the above lemmas to derive the main theorem for error bounds.

Proof of Theorem 3.1. Recall the definitions in Section 3.2,

$$\begin{aligned}\delta\mathcal{L}(\Delta_Z) :=&\mathcal{L}(Z^* + \Delta Z) - \mathcal{L}(Z^*) - \langle \nabla_Z \mathcal{L}(Z^*), X_1\Delta_\Theta\rangle - \langle \nabla_Z \mathcal{L}(Z^*), X_2\Delta_G\rangle,\\ F(\Delta_\Theta, \Delta_G) :=&\mathcal{L}(Z^* + \Delta_Z) - \mathcal{L}(Z^*)\\ &+ \lambda_1[\mathcal{R}_1(\Theta^* + \Delta_\Theta) - \mathcal{R}_1(\Theta^*)] + \lambda_2[\mathcal{R}_2(G^* + \Delta_G) - \mathcal{R}_2(G^*)].\end{aligned}$$

Therefore,

$$\begin{aligned}F(\Delta_\Theta, \Delta_G) =&\delta L(\Delta_Z) + \langle \nabla_Z \mathcal{L}(Z^*), X_1\Delta_\Theta\rangle + \langle \nabla_Z \mathcal{L}(Z^*), X_2\Delta_G\rangle\\ &+ \lambda_1[\mathcal{R}_1(\Theta^* + \Delta_\Theta) - \mathcal{R}_1(\Theta^*)] + \lambda_2[\mathcal{R}_2(G^* + \Delta_G) - \mathcal{R}_2(G^*)]\\ =&\delta\mathcal{L} - 2\langle X_1^T(Y - X_1\Theta^* - X_2G^*), \Delta_\Theta\rangle - 2\langle X_2^T(Y - X_1\Theta^* - X_2G^*), \Delta_G\rangle\\ &+ \lambda_1[\mathcal{R}_1(\Theta^* + \Delta_\Theta) - \mathcal{R}_1(\Theta^*)] + \lambda_2[\mathcal{R}_2(G^* + \Delta_G) - \mathcal{R}_2(G^*)]\\ :=&V_1 + V_2 + V_3 + V_4,\end{aligned}$$

where

$$\begin{aligned}V_1 =&\delta\mathcal{L},\\ V_2 =&-2\langle X_1^T(Y - X_1\Theta^* - X_2G^*), \Delta_\Theta\rangle - 2\langle X_2^T(Y - X_1\Theta^* - X_2G^*), \Delta_G\rangle,\\ V_3 =&\lambda_1[\mathcal{R}_1(\Theta^* + \Delta_\Theta) - \mathcal{R}_1(\Theta^*)],\\ V_4 =&\lambda_2[\mathcal{R}_2(G^* + \Delta_G) - \mathcal{R}_2(G^*)].\end{aligned}$$

By Lemma 3.5

$$V_1 \geq \overline{\mathcal{K}}(\|\Delta_\Theta\|_F^2 + \|\Delta_G\|_F^2) - \overline{\tau}(Z^*). \tag{3.14}$$

By the inequality (i) and (ii) in equation (3.11), we obtain

$$\begin{aligned}V_2 =&-2\langle X_1^T(Y - X_1\Theta^* - X_2G^*), \Delta_\Theta\rangle - 2\langle X_2^T(Y - X_1\Theta^* - X_2G^*), \Delta_G\rangle\\ \geq&-\mathcal{R}_1^*(X_1^T\nabla_Z\mathcal{L}(Z^*))\mathcal{R}_1(\Delta_\Theta) - \mathcal{R}_2^*(X_2^T\nabla_Z\mathcal{L}(Z^*))\mathcal{R}_2(\Delta_G),\\ \geq&-\frac{\lambda_1}{2}[\mathcal{R}_1(\Pi_{\overline{\mathcal{M}_1}}(\Delta_\Theta)) + \mathcal{R}_1(\Pi_{\overline{\mathcal{M}_1}^\perp}(\Delta_\Theta))]\\ &-\frac{\lambda_2}{2}[\mathcal{R}_2(\Pi_{\overline{\mathcal{M}_2}}(\Delta_G)) + \mathcal{R}_2(\Pi_{\overline{\mathcal{M}_2}^\perp}(\Delta_G))].\end{aligned} \tag{3.15}$$

By Lemma 3.2

$$\begin{aligned} V_3 =& \lambda_1[\mathcal{R}_1(\Theta^* + \Delta_\Theta) - \mathcal{R}_1(\Theta^*)], \qquad (3.16)\\ \geq& \lambda_1(\mathcal{R}_1(\Pi_{\overline{\mathcal{M}}_1^\perp}(\Delta_\Theta)) - \mathcal{R}_1(\Pi_{\overline{\mathcal{M}}_1}(\Delta_\Theta)) - 2\mathcal{R}_1(\Pi_{\overline{\mathcal{M}}_1^\perp}\Theta^*)),\\ V_4 =& \lambda_2[\mathcal{R}_2(G^* + \Delta_G) - \mathcal{R}_2(G^*)]\\ \geq& \lambda_2(\mathcal{R}_2(\Pi_{\overline{\mathcal{M}}_2^\perp}(\Delta_G)) - \mathcal{R}_2(\Pi_{\overline{\mathcal{M}}_2}(\Delta_G)) - 2\mathcal{R}_2(\Pi_{\overline{\mathcal{M}}_2^\perp}G^*)). \qquad (3.17) \end{aligned}$$

Combining the inequalities (3.14), (3.15), (3.16) and (3.17), and dropping the positive term $\mathcal{R}_1(\Pi_{\overline{\mathcal{M}}_1^\perp}(\Delta_\Theta))$ and $\mathcal{R}_2(\Pi_{\overline{\mathcal{M}}_2^\perp}(\Delta_G))$, we obtain:

$$\begin{aligned} & F(\Delta_\Theta, \Delta_G)\\ \geq& \overline{\mathcal{K}}\|\Delta_\Theta\|_F^2 - \frac{\lambda_1}{2}(3\mathcal{R}_1(\Pi_{\overline{\mathcal{M}}_1}(\Delta_\Theta)) + 4\mathcal{R}_1(\Pi_{\overline{\mathcal{M}}_1^\perp}\Theta^*))\\ +& \overline{\mathcal{K}}\|\hat{\Delta}_G\|_F^2 - \frac{\lambda_2}{2}(3\mathcal{R}_2(\Pi_{\overline{\mathcal{M}}_2}(\Delta_G)) + 4\mathcal{R}_2(\Pi_{\overline{\mathcal{M}}_2^\perp}G^*)) - \overline{\tau}(Z^*)\\ \geq& \overline{\mathcal{K}}\|\Delta_\Theta\|_F^2 - \frac{3\lambda_1}{2}\mathcal{R}_1(\Pi_{\overline{\mathcal{M}}_1}(\Delta_\Theta)) + \overline{\mathcal{K}}\|\Delta_G\|_F^2 - \frac{3\lambda_2}{2}\mathcal{R}_2(\Pi_{\overline{\mathcal{M}}_2}(\Delta_G))\\ & - \overline{\tau}(Z^*) - 2\lambda_1\mathcal{R}_1(\Pi_{\overline{\mathcal{M}}_1^\perp}\Theta^*) - 2\lambda_2\mathcal{R}_2(\Pi_{\overline{\mathcal{M}}_2^\perp}G^*). \end{aligned}$$

Using the definition of subspace compatibility constant, we get

$$\begin{aligned} F(\Delta_\Theta, \Delta_G) =& V_1 + V_2 + V_3 + V_4\\ \geq& \overline{\mathcal{K}}\|\Delta_\Theta\|_F^2 - \frac{3\lambda_1}{2}\Psi_1(\overline{\mathcal{M}}_1)\mathcal{R}_1(\Delta_\Theta) + \overline{\mathcal{K}}\|\Delta_G\|_F^2 - \frac{3\lambda_2}{2}\Psi_2(\overline{\mathcal{M}}_2)\mathcal{R}_2(\Delta_G)\\ & - \overline{\tau}(Z^*) - 2\lambda_1\mathcal{R}_1(\Pi_{\overline{\mathcal{M}}_1^\perp}\Theta^*) - 2\lambda_2\mathcal{R}_2(\Pi_{\overline{\mathcal{M}}_2^\perp}G^*). \end{aligned}$$

Denote $\Phi = \max\{\lambda_1\Psi_1(\overline{\mathcal{M}}_1), \lambda_2\Psi_2(\overline{\mathcal{M}}_2)\}$ and $2D = \overline{\tau}(Z^*) + 2\lambda_1\mathcal{R}_1(\Pi_{\overline{\mathcal{M}}_1^\perp}\Theta^*) + 2\lambda_2\mathcal{R}_2(\Pi_{\overline{\mathcal{M}}_2^\perp}G^*)$. Then,

$$F(\Delta_\Theta, \Delta_G) \geq \overline{\mathcal{K}}\|\Delta_\Theta\|_F^2 - \frac{3}{2}\Phi\mathcal{R}_1(\Delta_\Theta) + \overline{\mathcal{K}}\|\Delta_G\|_F^2 - \frac{3}{2}\Phi\mathcal{R}_2(\Delta_G) - 2D.$$

Using Lemma 3.6, we get:

$$F(\Delta_\Theta, \Delta_G) \geq \overline{\mathcal{K}}\frac{\delta^2}{2} - \frac{3}{2}\Phi\delta - 2D. \qquad (3.18)$$

Since $\overline{\mathcal{K}}\frac{\delta^2}{2} - \frac{3}{2}\Phi\delta - 2D > 0$ as long as $\delta > \frac{1.5\Phi+\sqrt{2.25\Phi^2+4\overline{\mathcal{K}}D}}{\overline{\mathcal{K}}}$, we can say that when $\delta > \frac{3\Phi+2\sqrt{\overline{\mathcal{K}}D}}{\overline{\mathcal{K}}} > \frac{1.5\Phi+\sqrt{2.25\Phi^2+4\overline{\mathcal{K}}D}}{\overline{\mathcal{K}}}$, $F(\Delta_\Theta, \Delta_G) > 0$.

Then using Lemma 3.4, we achieve the error bound:

$$||\hat{\Theta} - \Theta^*||_F + ||\hat{G} - G^*||_F \leq \frac{3\Phi + 2\sqrt{\overline{\mathcal{K}}D}}{\overline{\mathcal{K}}},$$

which completes the proof. □

Chapter 4

Investigation into Specific Settings

In chapter 3, we provided error bounds under four natural assumptions. In this chapter, we will investigate the four conditions for a specific model. More importantly, we establish an innovative proof for the structural incoherence property and provide specific results on the error bound.

4.1 A specific model with its four natural conditions

Let's consider a specific model with regularization functions $\mathcal{R}_1 = \|.\|_{1,2}$ and $\mathcal{R}_2 = \|.\|_1$. To be more specific,

$$Y = X_1\Theta^* + X_2G^* + W, \tag{4.1}$$

where $\Theta^* \in \mathbb{R}^{p\times q}$ is row-sparse, $G^* \in \mathbb{R}^{n\times q}$ is entry-wise sparse and $W \in \mathbb{R}^{n\times q}$ is the noise matrix whose Frobenius norm should be small. The M-estimator is

$$(\hat{\Theta}, \hat{G}) \in \arg\min_{\Theta,G}\{\|Y - X_1\Theta - X_2G\|_F^2 + \lambda_1\|\Theta\|_{1,2} + \lambda_2\|G\|_1\}. \tag{4.2}$$

Next, let's verify that the four conditions imposed on the regularization functions $\mathcal{R}_\alpha$ $(\alpha = 1, 2)$ and the loss function $\mathcal{L} = \|Y - X_1\Theta - X_2G\|_F^2$ are satisfied.

(C1) The loss function $\mathcal{L} = \|Y - X_1\Theta - X_2G\|_F^2$ is convex and differentiable.

(C2) The regularizers $\mathcal{R}_1 = \|.\|_{1,2}$ and $\mathcal{R}_2 = \|.\|_1$. are norms and are decomposable with respect to the subspace pairs $(\mathcal{M}_\alpha, \overline{\mathcal{M}}_\alpha^\perp)$, where $\mathcal{M}_\alpha \in \overline{\mathcal{M}}_\alpha$ $(\alpha = 1, 2)$.

Proposition 4.1. *The loss function and regularization functions in (4.2) satisfy the properties (C1) and (C2).*

Remark. *The proof is also omitted since it is very straightforward.*

The next condition is about the "restricted strong convexity". Since

$$\begin{aligned}\delta\mathcal{L}(\Delta_\Theta; \Theta^*, G^*) :=&\mathcal{L}(Z^* + X_1\Delta_\Theta) - \mathcal{L}(Z^*) - \langle \nabla_\Theta \mathcal{L}(Z^*), \Delta_\Theta\rangle \\ =&\|Y - Z^* - X_1\Delta_\Theta\|_F^2 - \|Y - Z^*\|_F^2 - \langle X_1^T \nabla_Z \mathcal{L}(Z^*), \Delta_\Theta\rangle \\ =&\|Y - X_1\Theta^* - X_2G^* - X_1\Delta_\Theta\|_F^2 - \|Y - X_1\Theta^* - X_2G^*\|_F^2 \\ &+ 2\langle X^T(Y - X_1\Theta^* - X_2G^*), \Delta_\Theta\rangle \\ =&\|X_1\Delta_\Theta\|_F^2, \end{aligned} \tag{4.3}$$

the "restricted strong convexity" reduces to $\|X_1\Delta_\Theta\|_F^2 \geq \mathcal{K}_L\|\Delta_\Theta\|_F^2 - \mathcal{G}\mathcal{R}_2(\Delta_\Theta)$.

(C3) [Restricted Strong Convexity]

$$\|X_1\Delta_\Theta\|_F^2 \quad \geq \mathcal{K}_L\|\Delta_\Theta\|_F^2 - \mathcal{G}_1\mathcal{R}_1(\Delta_\Theta), \tag{4.4}$$

$$\|X_2\Delta_G\|_F^2 \quad \geq \mathcal{K}_L\|\Delta_G\|_F^2 - \mathcal{G}_2\mathcal{R}_2(\Delta_G). \tag{4.5}$$

Proposition 4.2. *Model (4.1) satisfies the Restricted Strong Convexity.*

Actually, there is a similar result in [1], which proves

$$\frac{1}{n}\|X_1\Delta_\Theta\|_F^2 \geq \kappa_L\|\Delta_\Theta\|_F^2 - g_1\mathcal{R}_1(\Delta_\Theta), \tag{4.6}$$

where κ_L and g_1 are positive constants. We can prove this proposition with $\mathcal{K}_L = n\kappa_L, \mathcal{G}_1 = ng_1, \mathcal{G}_1 = ng_2$. The detail proof is omitted.

The next condition is the structural incoherence property. Since in this specific setting, $\mathcal{L} = \|Y - X_1\Theta - X_2G\|_F^2$, the structural incoherence condition reduces to the following form.

(C4) [Structural Incoherence]

$$2\langle X_1\Delta_\Theta, X_2\Delta_G\rangle \leq \frac{\mathcal{K}_L}{2}(\|\Delta_\Theta\|_F^2 + \|\Delta_G\|_F^2) + \sum_{\alpha=1,2} \mathcal{H}_\alpha \mathcal{R}_\alpha^2(\Delta_\alpha). \tag{4.7}$$

In the following context, we provide a theorem to guarantee the structural incoherence for the setting (4.2).

Theorem 4.1 (Structural Incoherence Theorem)**.** *Assume that X_2 is an $n \times n$ random matrix whose entries $a_{i,j}$ are independent random variables with fourth moment bounded by 1, and that X_1 is an $n \times p$ random matrix $X_1^T = B = (B_1, ..., B_n)$ with $\mathbb{E}\|X_1\|^2 \leq \frac{\mathcal{K}_L}{16\Lambda^2}$, where $\Lambda = \max \frac{2+3\lambda_{\gamma_1} * \Psi_{\gamma_1}(\overline{\mathcal{M}_{\gamma_1}})}{2+3\lambda_{\gamma_2} * \Psi_{\gamma_2}(\overline{\mathcal{M}_{\gamma_2}})}$, $\mathcal{K}_L = n\kappa_L$, and κ_L is the positive constant in formula (4.6). Assume the rows of X are independent and identically distributed. Then $|2\langle X_1\Delta\Theta, X_2\Delta G\rangle| \leq \frac{n\kappa_L}{2}\sum_\alpha \|\Delta_\alpha\|_F^2$.*

Actually, a bridge between $|2\langle X_1\Delta\Theta, X_2\Delta G\rangle|$ and $|\sigma_{\max}(\mathcal{P}_1^T X_1^T X_2 \mathcal{P}_2)|$ was established in [60], where $\mathcal{P}_\alpha$ $(\alpha = 1, 2)$ represents the projection operator on to the subspace $\mathcal{M}_\alpha$ $(\alpha = 1, 2)$. The structural incoherence is then related to the largest singular value of the product of two random matrices $\mathcal{P}_1^T X_1^T$ and $X_2\mathcal{P}_2$. In the following section, we study the largest singular value of the product of two random matrices. After that, we will prove the structural incoherence with $\mathcal{H}_\alpha = 0$ in Section 4.3.

4.2 Bound the largest singular values

The largest singular value is also called the spectral norm of a matrix, we denote it by $\|.\|$. In this section, we estimate the largest singular value of the product of two random matrices. The following theorem is our main result.

Theorem 4.2. *Let A be an $n \times n$ random matrix whose entries $a_{i,j}$ are independent random variables. Let B be an $p \times n$ random matrix $B = (B_1, ..., B_n)$, where B_i*

are independent and identically distributed. Then we can bound the largest singular value of matrix BA

$$\mathbb{E}\|BA\| \leq n\mathbb{E}a_{11}^2\mathbb{E}\|B\|^2 + C\log(2p)\frac{w_2 n}{w_1^3}, \tag{4.8}$$

where $w_1 = n\mathbb{E}a_1^2 \cdot \mathbb{E}\|B_1\|_2^2 + 2\binom{2}{n}(\mathbb{E}|a_1|)^2 \cdot (\mathbb{E}\|B_1\|_2)^2$ *and* $w_2 = n\mathbb{E}a_1^4\mathbb{E}\|B_1\|_2^4 + 2(\mathbb{E}a_1^2)^2\mathbb{E}(\|B\|^2 \cdot \|B\|_F^2)$.

In proving Theorem 4.2, a foundation step is the M. Rudelson's Theorem [43]. Here we present the theorem as a reference.

Theorem 4.3 (M. Rudelson)**.** *Let* $u_1, ..., u_p$ *be vectors in* $\mathbb{R}^n$*. Then, for every p≥1, one has*

$$(\mathbb{E}\|\sum_{i=1}^{p}\epsilon_i u_i \otimes u_i\|^p)^{1/p} \leq C(\sqrt{p} + \sqrt{\log n}) \cdot \max_i \|u_i\|_2 \cdot \|\sum_{i=1}^{p} u_i \otimes u_i\|^{1/2}. \tag{4.9}$$

In particular, for every $t > 0$*, with probability at least* $1 - 2ne^{-ct^2}$*, one has*

$$\|\sum_{i=1}^{p}\epsilon_i u_i \otimes u_i\| \leq t \cdot \max_i \|u_i\|_2 \cdot \|\sum_{i=1}^{p} u_i \otimes u_i\|^{1/2}. \tag{4.10}$$

Remark. *In this paper we use* C, c *to represent constants.*

The next lemma is a consequence of M. Rudelson's Theorem and a standard symmetrization argument.

Lemma 4.4. *Let* $X_1, ..., X_p$ *be independent random vectors in* $\mathbb{R}^n$ *such that*

$$\|\mathbb{E}X_j \otimes X_j\| \leq s \quad \textit{for every } j. \tag{4.11}$$

Then

$$\mathbb{E}\|\sum_{j=1}^{p} X_j \otimes X_j\| \leq ps + C\log(2n) \cdot \mathbb{E}\max_j \|X_j\|_2^2. \tag{4.12}$$

Proof. Let $\epsilon_1, ..., \epsilon_n$ be independent symmetric Bernoulli random variables. Then by the triangle inequality, the standard symmetrization argument and the assumption

(4.11), we obtain

$$E := \mathbb{E}\| \sum_{j=1}^{p} X_j \otimes X_j \| \leq \mathbb{E}\| \sum_{j=1}^{p} (X_j \otimes X_j - \mathbb{E}X_j \otimes X_j)\| + \| \sum_{j=1}^{p} \mathbb{X}_j \otimes X_j \|$$
$$\leq 2\mathbb{E}\| \sum_{j=1}^{p} \epsilon_j X_j \otimes X_j \| + ps. \tag{4.13}$$

Condition on the random variables $X_1, ..., X_p$ and use $\mathbb{E}_\epsilon$ to denote the conditional expectation with respect to $\epsilon_1, ...\epsilon_p$. Applying Theorem 4.3, we obtain

$$\mathbb{E}_\epsilon \| \sum_{j=1}^{p} \epsilon_j X_j \otimes X_j \| \leq C\sqrt{\log{(2n)}} \cdot \mathbb{E} \max_j \|X_j\|_2 \cdot \| \sum_{j=1}^{n} X_j \otimes X_j \|^{1/2}.$$

Take expectation with respect to $X_1, ..., X_n$, and apply Cauchy-Schwarz inequality. Then we get

$$E \leq C\sqrt{\log{(2n)}} \cdot (\mathbb{E} \max_j \|X_j\|_2^2)^{1/2} \cdot E^{1/2} + ps.$$

Therefore,

$$\mathbb{E}\| \sum_{j=1}^{n} X_j \otimes X_j \| \leq ps + C \log{(2n)} \cdot \mathbb{E} \max_j \|X_j\|_2^2.$$

□

Remark. *The standard symmetrization is a popular technique in dealing with random matrices. Let matrix $A = (a_{ij})$, and let A' be an independent copy of A. Let ϵ_{ij} be independent symmetric Bernoulli random variables. Then by Jensen's inequality, $\mathbb{E}\|BA\| = \mathbb{E}\|B(A - \mathbb{E}A')\| \leq \mathbb{E}\|B(A - A')\| = \mathbb{E}\|B(\epsilon_{ij}(a_{ij} - a'_{ij}))\| \leq 2\mathbb{E}\|B(\epsilon_{ij}a_{ij})\|$.*

To complete the proof of Theorem 4.2, we still need to present another two auxiliary lemmas.

Lemma 4.5. *Let $a_1, ..., a_n$ be independent random variables. Let B be an $p \times n$ matrix $B = (B_1, ...B_i, ...B_n)$, where B_i, the columns of the matrix B, are independent and identically distributed. Consider the random vector $X \in \mathbb{R}^p$ defined as*

$$X = \sum_{i=1}^{n} a_i B_i.$$

Then

$$\mathbb{E}\|X\|_2^2 \quad \leq n\mathbb{E}a_1^2 \cdot \mathbb{E}\|B_1\|_2^2 + 2\tbinom{2}{n}(\mathbb{E}|a_1|)^2 \cdot (\mathbb{E}\|B_1\|_2)^2, \tag{4.14}$$

$$Var(\|X\|_2^2) \quad \leq n\mathbb{E}a_1^4\mathbb{E}\|B_1\|_2^4 + 2(\mathbb{E}a_1^2)^2\mathbb{E}(\|B\|^2 \cdot \|B\|_F^2). \tag{4.15}$$

Proof. To prove the inequality (4.14), we rewrite $\mathbb{E}\|X\|_2^2$ and use the properties of norms. Specificly,

$$\begin{aligned}
\mathbb{E}\|X\|_2^2 =&\mathbb{E}\|\sum_{i=1}^n a_iB_i\|_2^2\\
=&\mathbb{E}\|a_1B_1 + a_2B_2 + ... + a_nB_n\|_2^2\\
=&\mathbb{E}(\|a_1B_1 + a_2B_2 + ... + a_nB_n\|_2)^2\\
\leq&\mathbb{E}(\|a_1B_1\|_2 + \|a_2B_2\|_2 + ... + \|a_NB_n\|_2)^2\\
=&\mathbb{E}(\|a_1B_1\|_2^2 + \|a_2B_2\|_2^2 + ... + \|a_nB_n\|_2^2 + 2\sum_{1\leq i<j\leq n}\|a_iB_i\|_2 \cdot \|a_jB_j\|_2)\\
=&\mathbb{E}(a_1^2\|B_1\|_2^2 + a_2^2\|B_2\|_2^2 + ... + a_n^2\|B_n\|_2^2 + 2\sum_{1\leq i<j\leq n}|a_i| \cdot \|B_i\|_2 \cdot |a_j| \cdot \|B_j\|_2)\\
=&\mathbb{E}a_1^2\|B_1\|_2^2 + \mathbb{E}a_2^2\|B_2\|_2^2 + ... + \mathbb{E}a_n^2\|B_n\|_2^2 + 2\mathbb{E}\sum_{1\leq i<j\leq n}|a_i| \cdot |a_j| \cdot \|B_i\|_2 \cdot \|B_j\|_2\\
=&n\mathbb{E}(a_1^2\|B_1\|_2^2) + 2\sum_{1\leq i<j\leq n}\mathbb{E}(|a_i| \cdot |a_j| \cdot \|B_i\|_2 \cdot \|B_j\|_2)\\
=&n\mathbb{E}a_1^2 \cdot \mathbb{E}\|B_1\|_2^2 + 2\sum_{1\leq i<j\leq n}\mathbb{E}|a_i| \cdot \mathbb{E}|a_j| \cdot \mathbb{E}\|B_i\|_2 \cdot \mathbb{E}\|B_j\|_2\\
=&n\mathbb{E}a_1^2 \cdot \mathbb{E}\|B_1\|_2^2 + 2\sum_{1\leq i<j\leq n}\mathbb{E}|a_i| \cdot \mathbb{E}|a_j| \cdot \mathbb{E}\|B_i\|_2 \cdot \mathbb{E}\|B_j\|_2\\
=&n\mathbb{E}a_1^2 \cdot \mathbb{E}\|B_1\|_2^2 + 2\binom{2}{n}(\mathbb{E}|a_1|)^2 \cdot (\mathbb{E}\|B_1\|_2)^2,
\end{aligned}$$

which concludes the proof for (4.10).

Next we derive the upper bound for $Var(\|X\|_2^2)$. Since $Var(\|X\|_2^2)$ can be written as $Var(\|X\|_2^2) = \mathbb{E}\|X\|_2^4 - (\|X\|_2^2)^2$, we estimate $\mathbb{E}\|X\|_2^4$ and $(\|X\|_2^2)^2$ separately and then combine them. For $\mathbb{E}\|X\|_2^4$, we have

$$\begin{aligned}
\mathbb{E}\|X\|_2^4 = &\quad \mathbb{E}\langle X, X\rangle^2 = \mathbb{E}\langle \textstyle\sum_{i=1}^n a_iB_i, \sum_{j=1}^n a_jB_j\rangle^2\\
= &\quad \textstyle\sum_{i,j,k,l=1}\mathbb{E}a_ia_ja_ka_l\langle B_i, B_j\rangle\langle B_k, B_l\rangle
\end{aligned} \tag{4.16}$$

Due to the independence and mean zero assumption, non zero terms can only be of the following cases: $i = j = k = l; i = j, k = l, i \neq k; i = k, j = l, i \neq j; i = l, j = k, i \neq j$. Then the formula (4.16) is reduced to

$$\begin{aligned}
&\mathbb{E}\|X\|_2^4 \\
=&\sum_{i=1}^{n} \mathbb{E}a_i^4 \langle B_i, B_i\rangle^2 + \sum_{i,k=1,i\neq k}^{n} \mathbb{E}a_i^2 a_k^2 \langle B_i, B_i\rangle\langle B_k, B_k\rangle + 2\sum_{i,j=1,i\neq j}^{n} \mathbb{E}a_i^2 a_j^2 \langle B_i, B_j\rangle^2 \\
=&\sum_{i=1}^{n} \mathbb{E}a_i^4 \mathbb{E}\langle B_i, B_i\rangle^2 + \sum_{i,k=1,i\neq k}^{n} \mathbb{E}a_i^2 \mathbb{E}a_k^2 \mathbb{E}\langle B_i, B_i\rangle\mathbb{E}\langle B_k, B_k\rangle + 2\sum_{i,j=1,i\neq j}^{n} \mathbb{E}a_i^2 \mathbb{E}a_j^2 \mathbb{E}\langle B_i, B_j\rangle^2 \\
=&n\mathbb{E}a_1^4 \mathbb{E}\|B_1\|_2^4 + \sum_{i,k=1,i\neq k}^{n} \mathbb{E}a_i^2 \mathbb{E}a_k^2 \mathbb{E}\langle B_i, B_i\rangle\mathbb{E}\langle B_k, B_k\rangle + 2\sum_{i,j=1,i\neq j}^{n} \mathbb{E}a_i^2 \mathbb{E}a_j^2 \mathbb{E}\langle B_i, B_j\rangle^2 \\
=&n\mathbb{E}a_1^4 \mathbb{E}\|B_1\|_2^4 + \sum_{i,k=1,i\neq k}^{n} \mathbb{E}a_i^2 \mathbb{E}a_k^2 \mathbb{E}\langle B_i, B_i\rangle\mathbb{E}\langle B_k, B_k\rangle + 2(\mathbb{E}a_1^2)^2 \mathbb{E}\sum_{i,j=1,i\neq j}^{n} \langle B_i, B_j\rangle^2 \\
=&:I_1 + I_2 + I_3,
\end{aligned}$$

where

$$\begin{aligned}
I_1 =& n\mathbb{E}a_1^4 \mathbb{E}\|B_1\|_2^4, \\
I_2 =& \sum_{i,k=1,i\neq k}^{n} \mathbb{E}a_i^2 \mathbb{E}a_k^2 \mathbb{E}\langle B_i, B_i\rangle\mathbb{E}\langle B_k, B_k\rangle \leq (\mathbb{E}\|X\|_2^2)^2, \\
I_3 =& 2(\mathbb{E}a_1^2)^2 \mathbb{E}\sum_{i,j=1,i\neq j}^{n} \langle B_i, B_j\rangle^2 \\
=& 2(\mathbb{E}a_1^2)^2 \mathbb{E}\|B^* B\|_F^2 \\
=& 2(\mathbb{E}a_1^2)^2 \mathbb{E}(\|B\|^2 \|B\|_F^2).
\end{aligned}$$

Therefore

$$\begin{aligned}
Var(\|X\|_2^2) = \mathbb{E}\|X\|_2^4 - (\|X\|_2^2)^2 &= I_1 + I_2 + I_3 - (\|X\|_2^2)^2 \\
&\leq I_1 + I_3 \leq n\mathbb{E}a_1^4 \mathbb{E}\|B_1\|_2^4 + 2(\mathbb{E}a_1^2)^2 \mathbb{E}(\|B\|^2 \|B\|_F^2) \qquad (4.17)
\end{aligned}$$

□

Lemma 4.6. *Let A be an $n \times n$ random matrix whose entries $a_{i,j}$ are independent random variables. Let B be an $p \times n$ random matrix $B = (B_1, ..., B_n)$, where the* *columns B_i are independent and identically distributed. Let $X_1, ..., X_n \in \mathbb{R}^p$ denote the columns of the matrix BA. Then*

$$\mathbb{E} \max_{j=1,...,n} \|X_j\|_2^2 \leq \frac{w_2 n}{w_1^3}, \tag{4.18}$$

where $w_1 = n\mathbb{E}a_1^2 \cdot \mathbb{E}\|B_1\|_2^2 + 2\binom{2}{n}(\mathbb{E}|a_1|)^2 \cdot (\mathbb{E}\|B_1\|_2)^2$ and $w_2 = n\mathbb{E}a_1^4\mathbb{E}\|B_1\|_2^4 + 2(\mathbb{E}a_1^2)^2\mathbb{E}(\|B\|^2 \cdot \|B\|_F^2)$.

Proof. Let $B = (B_1, ..., B_n), A = (a_{ij})$. Then $X_j = \sum_{i=1}^n B_i a_{ij},\ j = 1, ...n$. Fix $j \in \{1, ..., n\}$, by the previous lemma, we get

$$\mathbb{E}\|X_j\|_2^2 \leq w_1, Var(\|X_j\|_2^2) \leq w_2,$$

where

$$\begin{aligned} w_1 =& n\mathbb{E}a_1^2 \cdot \mathbb{E}\|B_1\|_2^2 + 2\binom{2}{n}(\mathbb{E}|a_1|)^2 \cdot (\mathbb{E}\|B_1\|_2)^2, \\ w_2 =& n\mathbb{E}a_1^4\mathbb{E}\|B_1\|_2^4 + 2(\mathbb{E}a_1^2)^2\mathbb{E}(\|B\|^2 \cdot \|B\|_F^2). \end{aligned}$$

Recall Chebychev's inequality, which states that if Z is a random variable with $\sigma^2 = Var(Z)$. Then for arbitrary $k > 0$, $\mathbb{P}(|Z - \mathbb{E}Z| > k\sigma) \leq 1/k^2$.
Applying Chebychev's inequality for $Z = \|X_j\|_2^2,\ k = t\sqrt{w_2}$ where $t > 0$ is arbitrary, one can obtain

$$\mathbb{P}(|\|X_j\|_2^2 - \mathbb{E}\|X_j\|_2^2| > k\sigma) \leq 1/k^2. \tag{4.19}$$

Then

$$\begin{aligned} 1/k^2 \geq& \mathbb{P}(|\|X_j\|_2^2 - \mathbb{E}\|X_j\|_2^2| > k\sigma) \geq \mathbb{P}(\|X_j\|_2^2 > \mathbb{E}\|X_j\|_2^2| + k\sigma) \\ \geq& \mathbb{P}(\|X_j\|_2^2 > w_1 + k\sqrt{w_2}). \end{aligned} \tag{4.20}$$

Since $k = t\sqrt{w_2}$, $\frac{1}{t^2 w_2^2} \geq \mathbb{P}(\|X_j\|_2^2 > w_1 + tw_2)$. Taking the union bound over all $j = 1, ...p$, we get $\mathbb{P}(\max_{j=1,...p} \|X_j\|_2^2 > w_1 + tw_2) \leq \frac{n}{t^2 w_2}$. Integration completes the proof. □

Equipped with the above lemmas, we can now present a complete proof for Theorem 4.2

Proof of Theorem 4.2. Let $X_1, ..., X_n \in \mathbb{R}^p$ denote the columns of matrix BA, i.e., $BA = (X_1, ..., X_n)$. Then we can apply Lemma 4.4 to achieve the bound. Let's first check the conditions in Lemma 4.4 .

$$X_j = \sum_{i=1}^{n} B_i a_{ij}, \quad j = 1, ...n. \tag{4.21}$$

Since $\mathbb{E}\|X_j \otimes X_j\| = \mathbb{E}\langle X_j, x\rangle^2$ for some arbitrary vector $x \in S^{p-1}$, we can get

$$\begin{aligned}\mathbb{E}\|X_j \otimes X_j\| =& \mathbb{E}\langle X_j, x\rangle^2 = \mathbb{E}\langle \sum_{i=1}^{n} a_{ij} B_i, x\rangle^2 = \mathbb{E}(\sum_{i=1}^{n} a_{ij}\langle B_i, x\rangle)^2 \\ =& \sum_{i=1}^{n} \mathbb{E}a_{ij}^2 \mathbb{E}\langle B_i, x\rangle^2 = \mathbb{E}a_{11}^2 \sum_{i=1}^{n} \mathbb{E}\langle B_i, x\rangle^2 \\ =& \mathbb{E}a_{11}^2 \mathbb{E} \sum_{i=1}^{n} \langle B_i, x\rangle^2 = \mathbb{E}a_{11}^2 \mathbb{E}\|B^* x\|_2^2 \\ \leq& \mathbb{E}a_{11}^2 \mathbb{E}\|B^*\|^2 = \mathbb{E}a_{11}^2 \mathbb{E}\|B\|^2.\end{aligned}$$

Applying Lemma 4.4, we obtain

$$\mathbb{E}\|BA\| = \mathbb{E}\|\sum_{j=1}^{p} X_j \otimes X_j\| \leq n\mathbb{E}a_{11}^2 \mathbb{E}\|B\|^2 + C\log(2p) \cdot \mathbb{E}\max_j \|X_j\|_2^2.$$

Then applying Lemma 4.6, we obtain

$$\mathbb{E}\|BA\| = \mathbb{E}\|\sum_{j=1}^{p} X_j \otimes X_j\| \leq n\mathbb{E}a_{11}^2 \mathbb{E}\|B\|^2 + C\log(2p)\frac{w_2 n}{w_1^3},$$

where $w_1 = n\mathbb{E}a_1^2 \cdot \mathbb{E}\|B_1\|_2^2 + 2\binom{2}{n}(\mathbb{E}|a_1|)^2 \cdot (\mathbb{E}\|B_1\|_2)^2$ and $w_2 = n\mathbb{E}a_1^4 \mathbb{E}\|B_1\|_2^4 + 2(\mathbb{E}a_1^2)^2 \mathbb{E}(\|B\|^2 \cdot \|B\|_F^2)$. □

4.3 Structural incoherence and error bounds

With the above lemmas and corollaries, we can derive a complete proof for the structural incoherence property.

Proof of Theorem 4.1. It has been shown in [60] that

$$|2\langle X_1\Delta\Theta, X_2\Delta G\rangle| \leq 2|\sigma_{\max}(\mathcal{P}_1^T X_1^T X_2 \mathcal{P}_2)| \times \Lambda^2 \times 2 \times (\|\Delta\alpha\|_F^2 + \|\Delta G\|_F^2)$$

where

$$\Lambda = \max \frac{2 + 3\lambda_{\gamma_1} * \Psi_{\gamma_1}(\overline{\mathcal{M}_{\gamma_1}})}{2 + 3\lambda_{\gamma_2} * \Psi_{\gamma_2}(\overline{\mathcal{M}_{\gamma_2}})}.$$

Apply Theorem 4.2 and do a simple computation. Then we obtain:

$$\mathbb{E}\|\mathcal{P}_2^T X_2^T X_1 \mathcal{P}\| \le n\mathbb{E}a_{11}^2 \mathbb{E}\|B\|^2 + C\log(2p)\frac{w_2 n}{w_1^3} \le \frac{n\kappa_L}{8\Lambda^2}, \tag{4.22}$$

where

$$\begin{aligned} w_1 =& n\mathbb{E}a_{11}^2 \cdot \mathbb{E}\|B_1\|_2^2 + 2\binom{2}{n}(\mathbb{E}|a_{11}|)^2 \cdot (\mathbb{E}\|B_1\|_2)^2, \\ w_2 =& n\mathbb{E}a_{11}^4 \mathbb{E}\|B_1\|_2^4 + 2(\mathbb{E}a_{11}^2)^2 \mathbb{E}(\|B\|^2 \cdot \|B\|_F^2), \\ \mathcal{K}_L =& n\kappa_L. \end{aligned}$$

Therefore, $|\sigma_{\max}(\mathcal{P}_1^T X_1^T X_2 \mathcal{P}_2)| \le \frac{\mathcal{K}_L}{8\Lambda^2}$.
Thus $|2\langle X_1\Delta\Theta, X_2\Delta G\rangle| \le \frac{1}{2}\mathcal{K}_L \sum_\alpha \|\Delta_\alpha\|_F^2$, where $\Lambda = \max \frac{2+3\lambda_{\gamma_1} * \Psi_{\gamma_1}(\overline{\mathcal{M}_{\gamma_1}})}{2+3\lambda_{\gamma_2} * \Psi_{\gamma_2}(\overline{\mathcal{M}_{\gamma_2}})}$, and $\mathcal{K}_L = n\kappa_L$. □

We have demonstrated the structural incoherence property for model (4.1). Next, we establish the error bound for it. In fact we will also do simulation for this model in Chapter 6,

Theorem 4.7. *Recall model (4.1) and its M-estimator (4.2). Assume* $\lambda_1 = 8n\sigma\sqrt{\frac{\log pq}{n}}$, *and* $\lambda_2 = 8n\sigma\{\sqrt{\frac{q}{n}} + \sqrt{\frac{\log p}{n}}\}$. *Then with high probability, the error of the the estimate* $(\hat{\Theta}, \hat{G})$ *is bounded by*

$$||\hat{\Theta} - \Theta^*||_F + ||\hat{G} - G^*||_F \le (\frac{24n\sigma}{\overline{\mathcal{K}}}) \max\{\sqrt{\frac{s(\log pq)}{n}}, \sqrt{s_r}(\sqrt{\frac{q}{n}} + \sqrt{\frac{\log p}{n}}\} \tag{4.23}$$

where

$$\begin{aligned} \overline{\mathcal{K}} =& \frac{\mathcal{K}_L}{2} - 64\overline{\mathcal{G}}^2\Phi^2, \\ \mathcal{K}_L =& n\kappa_L, \\ \overline{\mathcal{G}} =& \max_\alpha \frac{\sqrt{\mathcal{G}_\alpha + \mathcal{H}_\alpha}}{\lambda_\alpha}, \\ \Phi =& \max\{\lambda_1\Psi_1(\overline{\mathcal{M}_1}), \lambda_2\Psi_2(\overline{\mathcal{M}_2})\}, \end{aligned}$$

κ_L is the positive constant in(4.6), s_r is cardinality for the row-sparse matrix Θ, and s for the entry-sparse matrix G.

Proof. From Proposition 4.1 and Proposition 4.2, we know that conditions (C1)-(C3) are satisfied. Moreover, by Theorem 4.1, the structural incoherence is satisfied with high probability.

We know

$$2\mathcal{R}_1^*(\nabla_\Theta \mathcal{L}(\Theta^*, G^*)) = 2\mathcal{R}_1^*(X^T W) = 4 \max_{t=1,2,\dots p} \|(X^T W)_{t,*}\|_2 \leq 8n\sigma\{\sqrt{\frac{q}{n}} + \sqrt{\frac{\log p}{n}}\} = \lambda_1$$

with probability at least $1 - 2\exp(-2\log p)$, which extends the result in [60]. Also

$$2\mathcal{R}_2^*(\nabla_G \mathcal{L}(\Theta^*, G^*)) = 2\mathcal{R}_2^*(X^T W) = 4||(X^T W)_{*,t}||_\infty \leq 8n\sigma\sqrt{\frac{\log pq}{n}} = \lambda_2$$

with probability at least $1 - c_1\exp(-c_2(\log pq))$, which extends the result in [60]. Moreover, $\Psi_1(\overline{\mathcal{M}_1}) = \sum_\Delta \frac{||\Delta||_{1,2}}{||\Delta||_F} \leq \sqrt{s_r}$, and $\Psi_2(\overline{\mathcal{M}_2}) = \sup_\Delta \frac{||\Delta||_1}{||\Delta||_F} \leq \sqrt{s}$.
Applying Theorem 4.1, we get the error bound

$$||\hat{\Theta} - \Theta^*||_F + ||\hat{G} - G^*||_F \leq (\frac{24n\sigma}{\overline{\mathcal{K}}}) \max\{\sqrt{\frac{s(\log pq)}{n}}, \sqrt{s_r}(\sqrt{\frac{q}{n}} + \sqrt{\frac{\log p}{n}}\},$$

which completes the proof. □

4.4 Other examples

For the PCA model [1, 58, 60], $Y = \Theta^* + G^* + W$, where Θ^* is low-rank and G^* is element-wise sparse, the optimization problem is

$$\min_{\Theta,G}\{\|Y - \Theta - G\|_F^2 + \lambda_1\|\Theta\|_* + \lambda_2\|G\|_1\}. \tag{4.24}$$

We can specify our framework to this model by setting $X_2 = I$, and $X_1 = I$.

Another application is the multiple linear regression model [60] $Y = X(\Theta^* + G^*) + W$. The corresponding estimator is

$$(\hat{\Theta}, \hat{G}) \in \arg\min_{\Theta,G}\{\|Y - X\Theta - XG\|_F^2 + \lambda_1\|\Theta\|_{1,2} + \lambda_2\|G\|_1\}. \tag{4.25}$$

We can specify our framework to this model by setting $X_1 = X_2 = X$. For this case, the two matrices are greatly correlated, so we have to set more strict assumptions on the matrix to obtain structural incoherence. The following estimation of the largest singular values of matrix $X^T X$ was provided in [56].

Proposition 4.3. *Suppose that $X \in \mathbb{R}^{n \times n}$ is a Σ-Gaussian matrix. Then for any fixed i,k and every $\delta > 0$, with probability at least $1 - 4exp(-c_2 \delta^2)$, we have*

$$\|X^T X\| \leq n\|\Sigma\| + nc_1 max(\eta, \eta^2), \tag{4.26}$$

where $\eta = \sqrt{\frac{t}{n}} + \frac{\delta}{\sqrt{n}}$, and constants c_1, c_2 only depend on the distribution of the rows in X.

Chapter 5

Algorithm and Simulation

5.1 Algorithm

In this chapter, we do experiments for the specific model (4.1) where Θ^* is row-sparse and G^* is entry-sparse. The estimator is

$$(\hat{\Theta}, \hat{G}) \in \arg\min_{\Theta, G} \{\|Y - X_1\Theta - X_2 G\|_F^2 + \lambda_1 \|\Theta\|_{1,2} + \lambda_2 \|G\|_1\}. \tag{5.1}$$

For this model, we use FISTA (Fast Iterative Shrinkage-Thresholding Algorithm) [6] to solve the optimization problem.

In the first step, we approximate the objective function

$$F(\Theta, G) = \|Y - X_1\Theta - X_2 G\|_F^2 + \lambda_1 \|\Theta\|_{1,2} + \lambda_2 \|G\|_1 \tag{5.2}$$

with Q:

$$\begin{aligned} Q(\Theta, G) =& \|Y - X_1\Theta_{k-1} - X_2 G_{k-1}\|_F^2 - 2\langle X_1^T(Y - X_1\Theta_{k-1} - X_2 G_{k-1}), \Delta\Theta\rangle \\ & - 2\langle X_2^T(Y - X_1\Theta_{k-1} - X_2 G_{k-1}), \Delta G\rangle + \frac{L_1}{2}\|\Delta\Theta\|_F^2 + \frac{L_2}{2}\|\Delta G\|_F^2 \\ & + \lambda_1 \|\Theta\|_{1,2} + \lambda_2 \|G\|_1. \end{aligned}$$

Then, we calculate (Θ_k, G_k) using the formula Q.

$$\begin{aligned}
(\Theta_k, G_k) \in& \arg\min F(\Theta, G) \\
=& \arg\min\{\frac{L_1}{2}\|\Delta\Theta\|_F^2 + \frac{L_2}{2}\|\Delta G\|_F^2 - 2\langle X_1^T(Y - X_1\Theta_{k-1} - X_2G_{k-1}), \Delta\Theta\rangle \\
& - 2\langle X_2^T(Y - X_1\Theta_{k-1} - X_2G_{k-1}), \Delta G\rangle + \lambda_1\|\Theta\|_{1,2} + \lambda_2\|G\|_1\} \\
=& \arg\min\{\frac{L_1}{2}\|\Delta\Theta - \frac{1}{L_1}X_1^T(Y - X_1\Theta_{k-1} - X_2G_{k-1})\|_F^2 \\
& + \frac{L_2}{2}\|\Delta G - \frac{1}{L_2}X_2^T(Y - X_1\Theta_{k-1} - X_2G_{k-1})\|_F^2 + \lambda_1\|\Theta\|_{1,2} + \lambda_2\|G\|_1\} \\
=& \arg\min\{\frac{L_1}{2}\|\Theta - \Theta_{k-1} - \frac{1}{L_1}X_1^T(Y - X_1\Theta_{k-1} - X_2G_{k-1})\|_F^2 \\
& + \frac{L_2}{2}\|G - G_{k-1} - \frac{1}{L_2}X_2^T(Y - X_1\Theta_{k-1} - X_2G_{k-1})\|_F^2 + \lambda_1\|\Theta\|_{1,2} + \lambda_2\|G\|_1\}.
\end{aligned} \tag{5.3}$$

Actually formula (5.3) is in the form of

$$\arg\min\{\frac{L_1}{2}\|\Theta - A\|_F^2 + \frac{L_2}{2}\|G - B\|_F^2 + \lambda_1\|\Theta\|_{1,2} + \lambda_2\|G\|_1\}, \tag{5.4}$$

where

$$\begin{aligned}
A =& \Theta_{k-1} + \frac{1}{L_1}X_1^T(Y - X_1\Theta_{k-1} - X_2G_{k-1}), \\
B =& G_{k-1} + \frac{1}{L_2}X_2^T(Y - X_1\Theta_{k-1} - X_2G_{k-1}).
\end{aligned}$$

In fact, the iteration solution is

$$G_k = \tau_{\frac{\lambda_2}{L_2}}(B), \tag{5.5}$$

$$\Theta_{k(ij)} = \eta_i A_{ij}, \tag{5.6}$$

$$\eta_i = 1 - \frac{\lambda_1}{L_1\sqrt{A_{i1}^2 + A_{i2}^2 + ...A_{iq}^2}}, \tag{5.7}$$

where

$$\begin{aligned}
L_1 =& 2\lambda_{max}(X_1^T X_1),\\
L_2 =& 2\lambda_{max}(X_2^T X_2),\\
A =& \Theta_{k-1} + \frac{1}{L_1} X_1^T (Y - X_1\Theta_{k-1} - X_2 G_{k-1}),\\
B =& G_{k-1} + \frac{1}{L_2} X_2^T (Y - X_1\Theta_{k-1} - X_2 G_{k-1}).
\end{aligned}$$

Now we have obtained the necessary data for the FISTA simulation.

5.2 Simulation results

We conduct simulation using MATLAB to see the effect of structural incoherence which is the fundamental assumption of our model. In our experiments, we choose n=100. For the row-sparse matrix Θ^* with sub-Gaussian rows, we select nonzero rows randomly using MATLAB command 'randperm' and then generate row vectors from Gaussian distribution using MATLAB command 'randn'. The element-wise sparse matrix G^* is generated by the command 'sprand' in MATLAB.

We generate a random matrix X_2 which are to be used in both experiments. For the first set of experiments, we generate random matrix X_1 independently from X_2. In the second set of experiments, we set $X_1 = X_2$. Obviously, the first set of data enjoys better structural incoherence property. Then we study the effect of structural incoherence by observing the performances of the two groups. Moreover, we repeat the procedure according to different sparse levels. Thus we can see the effect of sparse levels in the performance of error bound. To be noted that, we carried out more than 20 tests for each situation, to average out the randomness and ensure the reliability of the experiments.

Refer to Figure 5.1 and Figure 5.2. We can see that the structural incoherence plays an important role in the performance of the errors. The more the structure is incoherent, the smaller is the error. Figure 5.1 shows that, the fewer nonzero rows in the original matrix Θ^*, the smaller is the error in the estimated Θ. Moreover, Figure 5.2 shows that, the fewer nonzero elements in the original matrix G^*, the smaller is the error in the estimated G. Those observations are consistent with our theory. Moreover, comparing the two figures, for the effect of structural incoherence, we observe much greater reduction in error in G, the element-sparse matrix. This performance is really caused by the property of structural incoherence. In fact, it indicates the great power of structural incoherence in estimating element-wise sparse matrix.

For the details of the MATLAB codes, please refer to the thesis package.

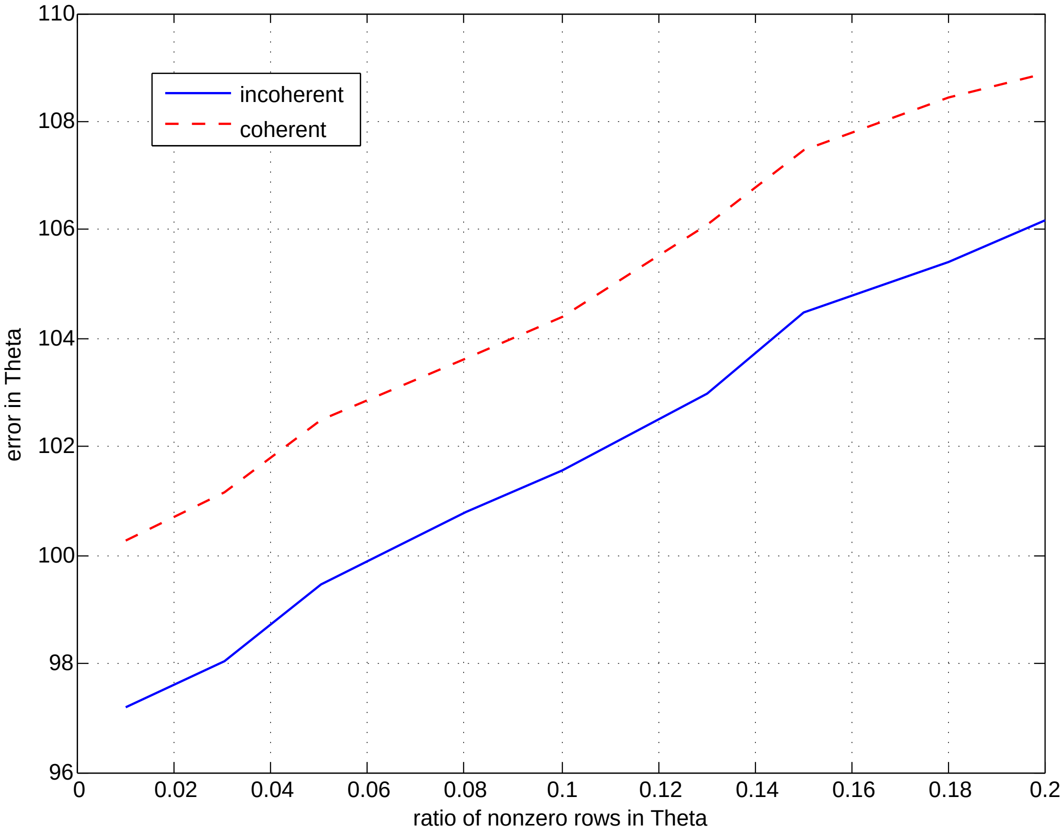

Figure 5.1: Performance of errors according to different sparse levels of Θ^*

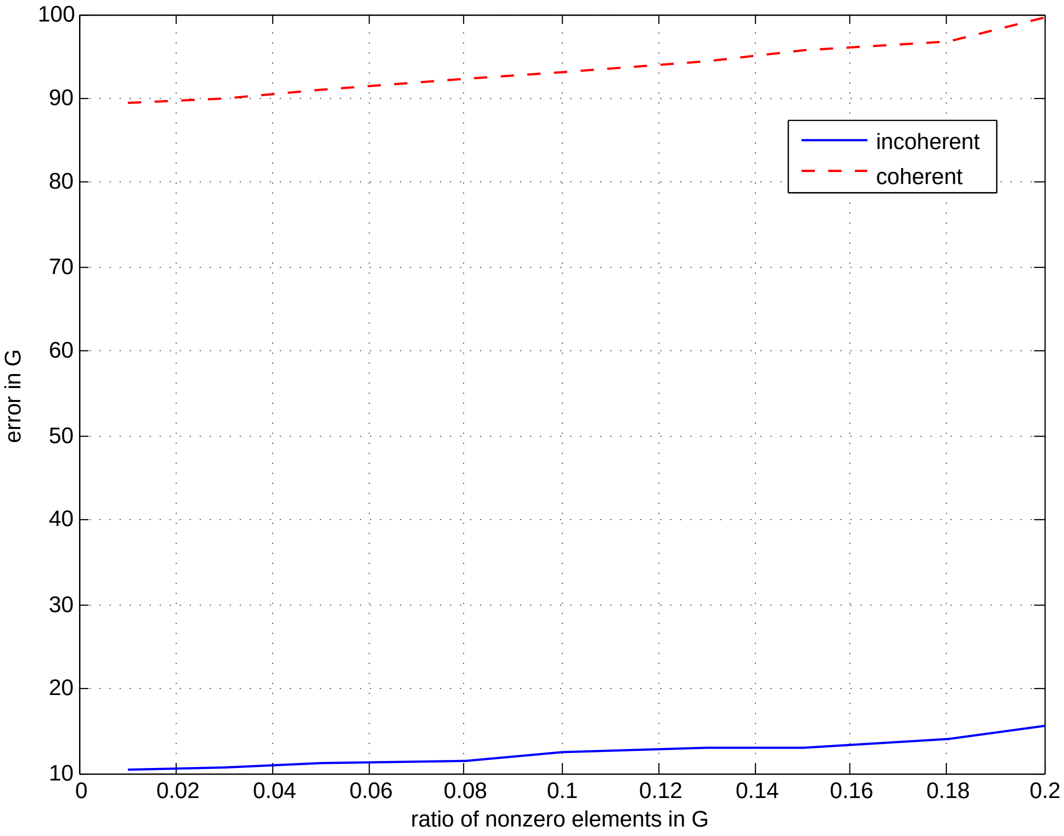

Figure 5.2: Performance of errors according to different sparse levels of G^*

Chapter 6

Conclusions

This thesis studied the structure decomposition problems in high-dimensional settings. We set up a general framework which involves distinct structures and imposed four natural assumptions on the model. Then we explored the four assumptions. In particular, we investigated the property of structural incoherence, and provided conditions under which the assumptions can hold in specific scenarios. The main results were the theoretical estimation on the error bound. And we then discussed structural incoherence for different specific scenarios, such as the PCA model and multi-regression model with gross errors. In the end, we conducted simulation to see the influence of the structural incoherence property. In fact, the simulation results provided good verifications for our theoretical analysis.

We should mention that the work done in this thesis is far from complete and comprehensive. There are still many interesting works to be done. Below we present some directions for further research that deserve more explorations.

- We only considered a number of scenarios that are special cases of our model. Maybe other norms and low-dimensional structures can also be incorporated into our framework under appropriate conditions.

- We only discussed two distinct structures in this paper. A future research direction is that, if the number of different structures is increased, whether one can possibly get some meaningful results concerning the parameter selection and the estimation of the error bound.

Bibliography

[1] A. Agarwal, S. Negahban, and M. J. Wainwright. Noisy matrix decomposition via convex relaxation: Optimal rates in high dimensions. *The Annals of Statistics*, 40(2):1171–1197, 2012.

[2] B. Schölkopf, J. Platt, and T. Hofmann. Multi-task feature learning. *Advances in Neural Information Processing Systems*, 19:41–48, 2006.

[3] F. R. Bach. Consistency of the group lasso and multiple kernel learning. *Journal of Machine Learning Research*, 9:1179–1225, 2008.

[4] F. R. Bach. Consistency of tracenorm minimization. *Journal of Machine Learning Research*, 9:1019–1048, June 2008.

[5] F. R. Bach, G. R. G. Lanckriet, and M. I. Jordan. Multiple kernel learning, conic duality, and the SMO algorithm. In *Proceedings of the 21st International Conference in Machine Learning*, New York, 2004. ACM.

[6] 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.

[7] P. J. Bickel, Y. Ritov , and A. Tsybakov. Simultaneous analysis of lasso and dantzig selector. *Annals of Statistics*, 35(6):2313–2351, 2007.

[8] P. J. Bickel and E. Levina. Covariance regularization by thresholding. *The Annals of Statistics*, 36(6):2577–2604, 2008.

[9] P. J. Bickel and E. Levina. Regularized estimation of large covariance matrices. *The Annals of Statistics*, 36(1):199–227, Feb 2008.

[10] F. Bunea, A. Tsybakov, and M. Wegkamp. Sparsity oracle inequalities for the lasso. *Electronic Journal of Statistics*, 1:169–194, 2007.

[11] T. T. Cai, C. H. Zhang, and H. H. Zhou. Optimal rates of convergence for sparse covariance matrix estimation. *The Annals of Statistics*, 38(4):2118–2144, 2010.

[12] E. Candès and T. Tao. Decoding by linear programming. *IEEE Transactions on Information Theory*, 51(12):4203–4215, December 2005.

[13] E. Candès and T. Tao. The dantzig selector: Statistical estimation when p is much larger than n. *Quality control and applied statistics*, 54(1):83–84, 2009.

[14] E. J. Candès, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *Journal of the ACM*, 58(3), May 2011.

[15] E. J. Candès and B. Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9(6):717–772, 2009.

[16] V. Cevher, P. Indyk, L. Carin, and R. G. Baraniuk. Sparse signal recovery and acquisition with graphical models. *Signal Processing Magazine, IEEE*, 27(6):92–103, Nov 2010.

[17] V. Chandrasekaran, P. A. Parrilo, and A. S. Willsky. Latent variable graphical model selection via convex optimization. *The Annals of Statistics*, 40(4):1935–1967, 2012.

[18] V. Chandrasekaran, S. Sanghavi, P. A. Parrilo, and A. S. Willsky. Rank-sparsity incoherence for matix decomposition. *SIAM Journal on Optimization*, 21(2), 2011.

[19] S. S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing*, 20(1):33–61, 1998.

[20] C. Zhang and J. Huang. Model selection consistency of the lasso selection in high-dimensional linear regression. *The Annals of Statistics*, pages 1567–1594, 2008.

[21] D. L. Donoho and J. Tanner. Neighborliness of randomly projected simplices in high dimensions. *Proc. Natl. Acad. Sci. USA*, 102(27):9452–9457, 2005.

[22] G. Obozinski and M. J. Wainwright and M. I. Jordan. Support union recovery in high-dimensional multivariate regression. *The Annals of Statistics*, 39(1):1–47, 2011.

[23] E. Greenshtein and Y. Ritov. Persistency in high dimensional linear predictor selection and the virtue of over-parametrization. *Bernoulli*, 10(6):971–988, 2004.

[24] D. Hsu, S. M. Kakade, and T. Zhang. Robust matrix decomposition with sparse corruptions. *IEEE Transactions on Information Theory*, 57:7221–7234, 2011.

[25] J. Huang and B. J. Frey. Cumulative distribution networks and the derivative-sum-product algorithm: Models and inference for cumulative distribution functions on graphs. *Journal of Machine Learning Research*, 12:301–348, 2011.

[26] J. Huang and T. Zhang. The benefit of group aparsity. *The Annals of Statistics*, 38(4):1978–2004, 2010.

[27] L. Jacob, G. Obozinski, and J. P. Vert. Group lasso with overlap and graph lasso. *the 26th International Conference on Machine Learning (ICML)*, pages 433–440, 2009.

[28] A. Jalali, P. Ravikumar, S. Sanghavi, and C. Ruan. A dirty model for multi-task learning. In *Advances in Neural Information Processing Systems 23 (NIPS 2010)*, 2010.

[29] R. Jenatton, J. Y. Audibert, and F. Bach. Active set algorithm for structured sparsity-inducing norms. In *2nd NIPS Workshop on Optimization for Machine Learning*, 2009.

[30] N. E. Karoui. Operator norm consistent estimation of large-dimensional sparse covariance matrices. *Annals of Statistics*, 36(6):2717–2756, 2008.

[31] R. H. Keshavan, A. Montanari, and S. Oh. Matrix completion from few entries. *IEEE Transactions on Information Theory*, 56(6):2980–2998, June 2010.

[32] R. H. Keshavan, A. Montanari, and S. Oh. Matrix completion from noisy entries. *Journal of Machine Learning Research*, 11:2057–2078, July 2010.

[33] Y. Kim, J. Kim, and Y. Kim. Blockwise sparse regression. (Sinica16375–390. MR2267240), 2006.

[34] V. Koltchinskii and M. Yuan. Sparse recovery in large ensembles of kernel machines. *In COLT'08*, 2008.

[35] C. Lam and J. Fan. Sparsistency and rates of convergence in large covariance matrix estimation. *Annals of Statistics*, 37:4254–4278, 2009.

[36] K. Lee and Y. Bresler. Guaranteed minimum rank approximation from linear observations by nuclear norm minimization with an ellipsoidal constraint. Technical report, UIUC, 2009.

[37] Z. Liu and L. Vandenberghe. Interior-point method for nuclear norm optimization with application to system identification. *SIAM Journal on Matrix Analysis and Applications*, 31(3):1235–1256, 2009.

[38] K. Lounici, A. B. Tsybakov, M. Pontil, and S. A. van de Geer. Taking advantage of sparsity in multi-task learning. In *22nd Conference On Learning Theory (COLT)*, 2009.

[39] M. Lustig, D. Donoho, J. Santos, and J. Pauly. Compressed sensing mri. *IEEE Signal Processing Magazine*, (27):72–82, March 2008.

[40] M. Mccoy and J. A. Tropp. Two proposals for robust pca using semidefinite programming. *Electronic Journal of Statistics*, 5:1123–1160, 2011.

[41] N. Meinshausen and P. Bühlmann. High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, 34:1436–1462, 2006.

[42] N. Meinshausen and B. Yu. Lasso-type recovery of sparse representations for high-dimensional data. *The Annals of Statistics*, 37(1):246–170, 2009.

[43] M. Rudelson. Random vectors in the isotropic positions. *Journal of Functional Analysis*, 164:60–72, 1999.

[44] Y. Nardi and A. Rinaldo. On the asymptotic properties of the group lasso estimator for linear models. *Electronic Journal of Statistics*, 2:605–633, 2008.

[45] S. Negahban, P. Ravikumar, M. J. Wainwright, and B. Yu. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Statistical Science*, 27(4):538–557, 2012.

[46] S. Negahban and M. J. Wainwright. Joint support recovery under high-dimensional scaling: Benefits and perils of l1,∞-regularization. In *Advances in Neural Information Processing Systems (NIPS)*, 2008.

[47] T. B. Obozinski, G. and M. I. Jordan. Joint covariate selection and joint subspace selection for multiple classification problems. *Statistics and Computing*, 20:231–252, 2010.

[48] A. Montanaria, R. H. Keshavan and S. Oh. Low-rank matrix completion with noisy observations: a quantitative comparison. In *Communication, Control, and Computing*, 2009.

[49] P. Ravikumar, M. J. Wainwright, G. Raskutti, and B. Yu. High-dimensional covariance estimation by minimizing $\ell1$-penalized log-determinant divergence. *Electron. J. Statist*, 5:935–980, 2011.

[50] B. Recht. A simpler approach to matrix completion. *Journal of Machine Learning Research*, 12:3413–3430, 2011.

[51] B. Recht, M. Fazel, and P. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Review*, 52(3):471–501, 2010.

[52] A. J. Rothman, P. J. Bickel, E. Levina, and J. Zhu. Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics*, 2:494–515, 2008.

[53] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58(1):267–288, 1996.

[54] B. Turlach, W. N. Venables, and S. J. Wright. Simultaneous variable selection. *Technometrics*, 47(3):349–363, 2005.

[55] S. A. van de Geer and P. Buhlmann. On the conditions used to prove oracle results for the lasso. *Electronic Journal of Statistics*, 3:1360–1392, 2009.

[56] R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. Technical report, Compressed Sensing: Theory and Applications, 2012.

[57] M. J. Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using $\ell1$-constrained quadratic programming (lasso). *IEEE Transactions on Information Theory*, 55(5):2183–2202, May 2009.

[58] H. Xu, C. Caramanis, and S. Sanghavi. Robust pca via outlier pursuit. Technical report, University of Texas, Austin, 2010.

[59] H. Xu and C. L. Leng. Robust multi-task regression with grossly corrupted observations. In *Proceedings of the 15th International Con- ference on Artificial Intelligence and Statistics (AISTATS)*, 2012.

[60] E. Yang and P. D. Ravikumar. Dirty statistical models. *NIPS*, pages 611–619, 2013.

[61] Y. Yuan, M. Lin. model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser.BStat. Methodol.*, 6849–67(MR2212574), 2006.

[62] H. Zhang, Y. Liu, Y. Wu, and J. Zhu. Variable selection for the multi-category svm via adaptive sup-norm regularization. *Electronic Journal of Statistics*, pages 1149–1167, 2008.

[63] C. H. Zhang and J. Huang. The sparsity and bias of the lasso selection in highdimensional linear regression. *Annals of Statistics*, 36(4):1567–1594, 2008.

[64] P. Zhao and B. Yu. On model selection consistency of lasso. *Journal of Machine Learning Research*, 7:2541–2567, 2006.

[65] P. Zhao, G. Rocha, and B. Yu. The composite absolute penalties family for grouped and hierarchical variable selection. *The Annals of Statistics*, 37(6A):3468–3497, 2009.

[66] S. Zhou, J. Lafferty, and L. Wasserman. Time-varying undirected graphs. In *In 21st Annual Conference on Learning Theory (COLT)*, Helsinki, Finland, July 2008.

[67] H. Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476), 2006.

A GENERAL FRAMEWORK FOR STRUCTURE DECOMPOSITION IN HIGH-DIMENSIONAL PROBLEMS

YANG JING

NATIONAL UNIVERSITY OF SINGAPORE

2014

A general framework for structure decomposition in high-dimensional problems **Yang Jing** **2014**