SOME METHODS FOR STATISTICAL INFERENCE USING HIGH-DIMENSIONAL LINEAR MODELS

by

TALAL AHMED

A dissertation submitted to the

School of Graduate Studies

Rutgers, The State University of New Jersey

In partial fulfillment of the requirements

For the degree of

Doctor of Philosophy

Graduate Program in Electrical and Computer Engineering

Written under the direction of

Waheed U. Bajwa

And approved by

_____

_____

_____

_____

New Brunswick, New Jersey

October, 2019

ABSTRACT OF THE DISSERTATION

Some Methods for Statistical Inference using High-dimensional Linear Models

By TALAL AHMED

Dissertation Director:

Waheed U. Bajwa

The ordinary linear model has been the bedrock of signal processing, statistics, and machine learning for decades. The last decade, however, has witnessed a marked transformation of this model: instead of the classical low-dimensional setting in which the sample size exceeds the number of features/predictors/variables, we are increasingly having to operate in the high-dimensional setting in which the number of variables far exceeds the sample size. Although such high-dimensional settings would ordinarily lead to ill-posed problems, the inference task has been studied under the rubric of high-dimensional statistical inference, where various notions of structure have been imposed on the model parameters to obtain unique solutions to the inference problem. While there are many statistical methods that guarantee unique solutions, these methods can easily become computationally prohibitive in ultrahigh-dimensional settings, in which the number of variables can scale exponentially with the sample size. In other cases, the traditional notions of structure on model parameters can be rather restrictive, especially when the variables naturally appear in the form of a multi-way array (tensor), as in the case of neuroimaging data analysis.

The purpose of this dissertation is to study inference using high-dimensional linear models for the cases when (i) the number of variables can scale exponentially with the number of samples, and (ii) the variables naturally form a tensor structure. Specifically, for each of these respective cases, the dissertation (i) proposes an efficient inference approach, (ii) provides high-probability performance guarantees for the proposed approach, and (iii) demonstrates efficacy of the inference approach in statistical analysis of real-world datasets.

Acknowledgements

It has been a privilege collaborating with my advisor Waheed Bajwa on a stream of exciting problems over the past few years. Over these years, he trained me in the art of thinking about problems on different levels of abstraction, while teaching me the value in mathematical rigor. I am grateful to my doctoral thesis committee members for their invaluable time and feedback: Emina Soljanin, Pierre Bellec, and Predrag Spasojevic. I am also grateful to my qualifying exam committee members for their challenging but intriguing discussions: Anand Sarwate, Kristin Dana, Salim El Rouayheb, and Vishal Patel.

During my time at Rutgers, I was lucky to be surrounded by the likes of Mudassir Shabbir, Zahra Shakeri, Mohsen Ghassemi, Tong Wu, with whom I not only shared a lot of fun moments but also engaged in fascinating discussions more number of times than I can possibly remember. During my summer internships away from Rutgers, I learnt immensely from collaborating with people from diverse academic backgrounds: Dagnachew Birru, Rittwik Jana, Zhibiao Rao, while finding wonderful friends in Arhum Savera, Deep Chakraborty, and Kat Poje.

Last, but not the least, I must mention the friend, who was by my side in all my ups and downs, with whom I shared the dankest as well as the most absurd moments of my graduate student life: Haroon Raja. Without his friendship, support, and uncanny wisdom, I probably won't have made it this far.

Table of Contents

Chapter 1

Introduction

1.1 Motivation

In this dissertation, we consider an ordinary linear model with response $y \in \mathbb{R}$, multidimensional predictors $\mathbf{X} \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d}$, multidimensional parameters $\mathbf{B} \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d}$, and noise $\eta \in \mathbb{R}$ such that $y = \langle \mathbf{X}, \mathbf{B} \rangle + \eta$, for $d \in \mathbb{Z}^+$. For the case of vector-valued predictors, i.e. $d = 1$, applications of this model can be found in genomics [1, 2] and text analysis [3, 4]. For the case of tensor-valued predictors, in which $d \geq 3$, applications of this model can be found in application areas like hyperspectral imaging [5–7] and neuroimaging [8, 9]. Define $\{\mathbf{X}_i\}_{i=1}^m$, $\{y_i\}_{i=1}^m$, and $\{\eta_i\}_{i=1}^m$ to be the realizations of $\mathbf{X}$, y, and η, respectively, where m refers to the number of samples/observations/measurements. Then, the realizations of the ordinary linear model can be expressed as

$$y_i = \langle \mathbf{X}_i, \mathbf{B} \rangle + \eta_i, \tag{1.1}$$

$i \in [[m]]$, where $[[a]] := \{1, 2, \ldots, a\}$ for any $a \in \mathbb{Z}^+$. Now, given $\{\mathbf{X}_i\}_{i=1}^m$ and $\{y_i\}_{i=1}^m$, we focus on the task of inference using the regression model in (1.1), which is equivalent to estimating/recovering/learning $\mathbf{B}$. Over the last decade, we are increasingly having to operate in the high-dimensional setting in which the number of variables is much greater than the sample size (i.e., $\prod_i n_i \gg m$). Ordinarily, such high-dimensional setting should lead to ill-posed problems. However, if some notion of structure can be imposed on $\mathbf{B}$ such as to constrain the *degrees of freedom* of the model, we *may* be able to formulate well-defined inference problems, even in the high-dimensional setting. Some important examples of such structure include tensor rank [10] and sparsity [11]. The focus of this dissertation is on studying the imposition of such structures that occur naturally in many data analysis problems, while characterizing the gains in computational or sample

complexity resulting because of the various structural assumptions.

1.2 Some Challenges in High-dimensional Inference

First, we consider inference using (1.1) for the case when $d = 1$. Specifically, we consider *ultrahigh-dimensional* linear models, in which the number of variables can scale exponentially with the sample size: $\log n_1 = \mathcal{O}(m^\alpha)$ for $\alpha \in (0, 1)$. The *principle of parsimony*—which states that only a small number of variables typically affect the response y—has been employed in the literature to help obtain unique solutions to inference problems based on high-dimensional linear models. There exist a number of techniques in the literature—such as forward selection/matching pursuit, backward elimination [12], least absolute shrinkage and selection operator (LASSO) [13], elastic net [14]—that can be employed for inference. However, all such existing inference techniques have super-linear (in the number of variables n_1) computational complexity. In the ultrahigh-dimensional setting, therefore, use of the traditional methods for statistical inference can easily become computationally prohibitive.

Second, we consider inference using (1.1) for the case when dealing when tensor data, i.e., $d \geq 3$. Tensor data appears naturally in areas like imaging and information sciences [15, 16], machine learning [17], signal processing [18], and quantum mechanics [19]. One simple approach for estimating parameter tensor $\mathbf{B}$ is to vectorize the tensors, and then use any of the traditional sparsity promoting methods for learning the regresison model. Specifically, the parameter tensor $\mathbf{B}$ and the predictor tensors $\{\mathbf{X}_i\}_{i=1}^m$ can be vectorized such that the model in (1.1) can equivalently be expressed as $y_i = \langle \text{vec}(\mathbf{X}_i), \text{vec}(\mathbf{B}) \rangle + \eta_i$, where vec(.) denotes the vectorization procedure. Given this vector-valued regression model, any of the aforementioned sparsity promoting techniques can be employed for estimating vec($\mathbf{B}$). However, a major drawback of vectorization is that the spatial structure among the entries of the tensor $\mathbf{B}$ is not preserved—structure that can possibly be exploited for efficient estimation of $\mathbf{B}$. To address this issue, various tensor decompositions have been considered with $\mathbf{B}$ to exploit the inter-modal relationships among the parameters.

Among the various tensor decompositions that capture such spatial relationships among tensor entries [16, 20], the notion of Tucker decomposition has been successfully employed to learn (1.1) under the imposition of low Tucker rank on $\mathbf{B}$ [10, 21, 22]. However, in the various convex and non-convex methods proposed for learning $\mathbf{B}$ under the imposition of low Tucker rank [21–23], the sample complexity of learning $\mathbf{B}$ has been shown to have linear scaling with n, where $n := \max\{n_i : i \in [[d]]\}$. Such sample complexity requirement can become prohibitive in application areas like neuroimaging data analysis [24]. Another challenge with the sole imposition of low Tucker rank on $\mathbf{B}$ is that the resulting regression model does not encompass the typical situation where the response depends on only a few of the (scalar) predictors in the model (i.e., the sparsity assumption).

1.3 Overview and Contributions

Our main contributions include (i) addressing the computational bottleneck for inference using ultra-high dimensional linear models by analyzing a two-step inference method, and (ii) addressing the sample complexity of inference using tensor-valued regression models by analyzing a new regression model and inference method. More specifically, our contributions are as follows:

1. Variable selection-based dimensionality reduction, commonly referred to as *variable screening*, has been put forth as a practical means of overcoming the computational bottleneck of inference using sparse high-dimensional linear models for $d = 1$. Since only a small number of (independent) variables actually contribute to the response (dependent variable) in the sparse setting, one can first—in principle—discard most of the variables (the screening step) and then carry out inference on a relatively low-dimensional linear model using any one of the sparsity-promoting techniques (the inference step). In this thesis, our focus is on obtaining understanding of the former step, i.e., the screening step. In Chapter 2, we revisit one of the simplest screening algorithms, which uses marginal correlations between

the variables $\{X_i\}_{i=1}^p$ and the response y for screening purposes [25, 26], and provide a comprehensive theoretical understanding of its screening performance for arbitrary ultrahigh-dimensional linear models. Our numerical experiments confirm that our new theoretical insights are not mere artifacts of analysis; rather, they are reflective of the challenges associated with marginal correlation-based variable screening.

2. In Chapter 3, we derive mathematical conditions for variable screening of two families of ultra-high dimensional linear models. The first family corresponds to sub-Gaussian linear models, in which the independent variables/predictors are independently drawn from (possibly different) sub-Gaussian distributions. The second family corresponds to arbitrary (random or deterministic) linear models in which the (empirical) correlations between independent variables satisfy certain polynomial-time verifiable conditions. The main result for this family of linear models establishes that, under appropriate conditions, it is possible to reduce the dimension of an ultrahigh-dimensional linear model to almost the sample size even when the number of active variables scales almost linearly with the sample size. This, to the best of our knowledge, is the first screening result that provides such explicit and optimistic guarantees *without* imposing a statistical prior on the distribution of the independent variables.

3. In Chapter 4, we study the tensor-valued regression model, i.e., the model in (1.1) for $d \geq 3$. In our work, we consider the simultaneous imposition of a certain low Tucker rank and sparse structure on the parameter tensor $\mathbf{B}$, massively reducing the degrees of freedom in $\mathbf{B}$. Subsequently, we formulate the estimation of $\mathbf{B}$ as a non-convex problem, and we propose a tensor variant of the projected gradient descent method to solve it. In contrast, prior works that study simultaneous imposition of multiple structures on tensor-valued regression models either (i) assume that the tensors satisfy certain cubic structures [27], or (ii) formulate a convex problem for estimating the parameter tensor [28], which can lead to sub-optimal sample complexity [23]. Furthermore, we provide theoretical analysis

to show the convergence behavior of the proposed algorithm, based on a certain Restricted Isometry Property. Finally, our experiments demonstrate the efficacy of the proposed method for learning the regression model in (1.1), under the simultaneous imposition of low rank and sparsity on $\mathbf{B}$.

4. In Chapter 5, we evaluate the Restricted Isometry Property for tensor-valued sub-Gaussian linear models, and in the process, we characterize the sample complexity of learning the posed tensor-valued regression model. Our sample complexity bound only has a polylogarithmic dependence on n, where $n := \max\{n_i : i \in [[d]]\}$. In contrast, prior works in tensor regression pose a sample complexity requirement that is either linear or super-linear in n [21–23]. In our experiments on real neuroimaging data, we demonstrate the utility of our proposed model and method in diagnosis of attention deficit hyperactivity disorder (ADHD) using fMRI images. Importantly, these experiments show that despite the imposition of low rank and sparse structure on the parameter tensor $\mathbf{B}$, and a massive reduction in degrees of freedom, our proposed model is not restrictive and is useful for neuroimaging data analysis.

1.4 Notation

Bold upper-case letters ($\mathbf{Z}$), upper-case letters (Z), bold lower-case letters ($\mathbf{z}$), lower-case letters (z), and underlined letters ($\underline{z}$) are used to denote tensors, matrices, vectors, scalars, and tuples, respectively. For any tuple $\underline{z}$ and scalar α, we use $\alpha\underline{z}$ to denote the tuple obtained by multiplying each entry of $\underline{z}$ by α. For any scalar $q \in \mathbb{Z}_+$, we use $[[q]]$ as a shorthand for $\{1, 2, \ldots, q\}$. Given $a \in \mathbb{R}$, $\lceil a \rceil$ denotes the smallest integer greater than or equal to a.

Given a vector $\mathbf{v}$, $\|\mathbf{v}\|_p$ denotes its ℓ_p norm, v_i denotes the i-th entry of $\mathbf{v}$, and $\mathbf{v}_{\min}$ denotes $\min_i |\mathbf{v}_i|$. Given two vectors $\mathbf{u} \in \mathbb{R}^n$ and $\mathbf{v} \in \mathbb{R}^n$ of same dimension, $\mathbf{u} \circ \mathbf{v}$ denotes the outer product, $\mathbf{u} \preceq \mathbf{v}$ denotes $u_i \leq v_i$ for all $i \in [[n]]$, $\mathbf{u} = \mathbf{v}$ denotes $u_i = v_i$ for all $i \in [[n]]$, and $\max\{\mathbf{u}, \mathbf{v}\}$ denotes entry-wise maxima.

Given a matrix U, U_j denotes its j-th column and $U_{i,j}$ denotes the entry in its i-th

row and j-th column, $U(:,i)$ denotes the i-th column, and $\|U\|_{1,2}$ denotes $\max_i \|U(:,i)\|_2$. Further, given a set $\mathcal{I} \subset \mathbb{Z}_+$, $U_{\mathcal{I}}$ (resp., $v_{\mathcal{I}}$) denotes a submatrix (resp., subvector) obtained by retaining columns of U (resp., entries of v) corresponding to the indices in $\mathcal{I}$. Finally, the superscript $(\cdot)^\top$ denotes the transpose operation.

Given any tensor $\mathbf{Z}$, the $(i_1, i_2, \ldots, i_d)$-th entry is given by $\mathbf{Z}(i_1, i_2, \ldots, i_d)$, the Frobenius norm $\|\mathbf{Z}\|_F$ is given by $\sqrt{\sum_{i_1,i_2,\ldots,i_d} \mathbf{Z}(i_1, i_2, \ldots, i_d)^2}$, the ℓ_1 norm is given by $\sum_{i_1,i_2,\ldots,i_d} |\mathbf{Z}(i_1, i_2, \ldots, i_d)|$, and the mode-$i$ matricization $\mathbf{Z}_{(i)}$ is the matrix obtained from column-arrangement of the mode-i fibers of $\mathbf{Z}$. The conjugate transpose of a given linear map, $\mathcal{X} : \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d} \to \mathbb{R}^m$, is denoted by $\mathcal{X}^* : \mathbb{R}^m \to \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d}$. Following the tensor notation in [16], for matrices $\widetilde{U}_i \in \mathbb{R}^{n_i \times r_i}$, $i \in [[d]]$, and tensor $\mathbf{S} \in \mathbb{R}^{r_1 \times r_2 \times \cdots \times r_d}$, we define $\mathbf{S} \times_1 \widetilde{U}_1 \times_2 \widetilde{U}_2 \cdots \times_d \widetilde{U}_d$ as $\sum_{i_1,i_2,\ldots,i_d} \mathbf{S}(i_1, i_2, \ldots, i_d)\, \widetilde{U}_1(:, i_1) \circ \widetilde{U}_2(:, i_2) \circ \cdots \circ \widetilde{U}_d(:, i_d)$.

Chapter 2

Correlation-based Variable Screening

Statistical inference can be computationally prohibitive in ultrahigh-dimensional linear models. Correlation-based variable screening, in which one leverages marginal correlations for removal of irrelevant variables from the model prior to statistical inference, can be used to overcome this challenge. Prior works on correlation-based variable screening either impose strong statistical priors on the linear model or assume specific post-screening inference methods. This chapter extends the analysis of correlation-based variable screening to arbitrary linear models and post-screening inference techniques. In particular, (i) it shows that a condition—termed the screening condition—is sufficient for successful correlation-based screening of linear models, and (ii) it provides insights into the dependence of marginal correlation-based screening on different problem parameters. Finally, numerical experiments confirm that the insights of this chapter are not mere artifacts of analysis; rather, they are reflective of the challenges associated with marginal correlation-based variable screening.

2.1 Introduction

In this chapter, our focus is on the ordinary linear model in (1.1) for the case when $d = 1$ (vector-valued predictors), and we define $n := n_1$ so that n denotes the number of features/predictors/variables in (1.1) for the case when $d = 1$. Further, for ease of notation in the case $d = 1$, we denote the ordinary linear model in (1.1) as $\mathbf{y} = X\boldsymbol{\beta}$+noise where the dimension, n, of $\boldsymbol{\beta}$ refers to number of variables; whereas, the dimension, m, of $\mathbf{y}$ refers to the sample size. Finally, we allow the number of variables, n, to scale exponentially with the number of samples, m, such that $\log n = \mathcal{O}(m^{\alpha})$ for $\alpha \in (0, 1)$ (ultrahigh-dimensional setting). In such ultra-high dimensional setting, it is reasonable

to hypothesize that only a small number of (independent) variables actually contribute to the response (dependent variable). While there exist a number of techniques in the literature—such as forward selection/matching pursuit and backward elimination [12], least absolute shrinkage and selection operator (LASSO) [13], elastic net [14], smoothly clipped absolute deviation (SCAD) [29], bridge regression [30, 31], adaptive LASSO [32], group LASSO [33], and Dantzig selector [34]—that can be employed for inference from high-dimensional linear models, all these techniques have super-linear (in the number of variables n) computational complexity, and thus these methods can quickly become computationally prohibitive in the ultrahigh-dimensional setting.

Variable selection-based dimensionality reduction, commonly referred to as *variable screening*, has been put forth as a practical means of overcoming this *curse of dimensionality* [35]: since only a small number of independent variables/predictors actually contribute to $\mathbf{y}$, one can first—in principle—discard most of the variables (the screening step) and then carry out inference on a relatively low-dimensional linear model using any one of the sparsity-promoting techniques (the inference step). There are two main challenges that arise in the context of variable screening in ultrahigh-dimensional linear models. First, the screening algorithm should have low computational complexity (ideally, $\mathcal{O}(mn)$). Second, the screening algorithm should be accompanied with mathematical guarantees that ensure the reduced linear model contains *all* relevant variables that affect the response. Our goal in this chapter is to revisit one of the simplest screening algorithms, which uses marginal correlations between the variables $\{X_i\}_{i=1}^n$ and the response $\mathbf{y}$ for screening purposes [25, 26], and provide a theoretical understanding of its screening performance for arbitrary ultrahigh-dimensional linear models.

2.1.1 Relationship to Prior Work

Researchers have long intuited that the (absolute) marginal correlation $|X_i^\top \mathbf{y}|$ is a strong indicator of whether the i-th variable contributes to the response variable. One of the earliest screening works in this regard that is agnostic to the choice of the subsequent inference techniques is termed *sure independence screening* (SIS) [36]. SIS is based on simple thresholding of marginal correlations and satisfies the so-called *sure screening*

property—which guarantees that all important variables survive the screening stage with high probability—for the case of normally distributed variables. An iterative variant of SIS, termed ISIS, is also discussed in [36], while [37] presents variants of SIS and ISIS that can lead to reduced false selection rates of the screening stage. Extensions of SIS to generalized linear models are discussed in [37, 38], while its generalizations for semi-parametric (Cox) models and non-parametric models are presented in [39, 40] and [41, 42], respectively.

The defining characteristics of the works referenced above is that they are agnostic to the inference technique that follows the screening stage. In recent years, screening methods have also been proposed for specific optimization-based inference techniques. To this end, [43] formulates a marginal correlations-based screening method, termed SAFE, for the LASSO problem and shows that SAFE results in zero false selection rate. In [44], the so-called *strong rules* for variable screening in LASSO-type problems are proposed that are still based on marginal correlations and that result in discarding of far more variables than the SAFE method. The screening tests of [43, 44] for the LASSO problem are further improved in [45–47] by analyzing the dual of the LASSO problem.

Notwithstanding these prior works, we have holes in our understanding of variable screening in ultrahigh-dimensional linear models. Works such as [43–47] necessitate the use of LASSO-type inference techniques after the screening stage. In addition, these works do not help us understand the relationship between the problem parameters and the dimensions of the reduced model. Similar to [36, 37, 48, 49], and in contrast to [43–47], our focus in this chapter is on screening that is agnostic to the post-screening inference technique. To this end, [48] lacks a rigorous theoretical understanding of variable screening using the generalized correlation. While [36, 37, 49] overcome this shortcoming of [48], these works have two major limitations. First, their results are derived under the assumption of restrictive statistical priors on the linear model (e.g., normally distributed X_i's). In many applications, however, it can be a challenge to ascertain the distribution of the independent variables. Second, the analyses in [36, 37, 49] assume the variance of the response variable to be bounded by a constant; this assumption, in turn, imposes the condition $\|\boldsymbol{\beta}\|_2 = \mathcal{O}(1)$. In contrast, defining

$\beta_{\min} := \min_i |\boldsymbol{\beta}_i|$, we establish in the sequel that the ratio $\frac{\beta_{\min}}{\|\boldsymbol{\beta}\|_2}$ (and not $\|\boldsymbol{\beta}\|_2$) directly influences the performance of marginal correlation-based screening procedures.

2.1.2 Our Contributions

Our focus in this chapter is on marginal correlation-based screening of high-dimensional linear models that is agnostic to the post-screening inference technique. To this end, we provide an extended analysis of the thresholding-based SIS procedure of [36]. The resulting screening procedure, which we term *extended sure independence screening* (ExSIS), provides new insights into marginal correlation-based screening of arbitrary high-dimensional linear models. Specifically, we first provide a simple, distribution-agnostic sufficient condition—termed the *screening condition*—for (marginal correlation-based) screening of linear models. This sufficient condition, which succinctly captures joint interactions among both the active and the inactive variables, is then leveraged to explicitly characterize the performance of ExSIS as a function of various problem parameters, including noise variance, the ratio $\frac{\beta_{\min}}{\|\boldsymbol{\beta}\|_2}$, and model sparsity. The numerical experiments reported at the end of this chapter confirm that the dependencies highlighted in this screening result are reflective of the actual challenges associated with marginal correlation-based screening and are not mere artifacts of our analysis.

The rest of this chapter is organized as follows. We formulate the problem of marginal correlation-based screening in Sec. 2.2. Next, in Sec. 2.3, we define the screening condition and present our main result that establishes the screening condition as a sufficient condition for successful variable screening. Finally, results of numerical experiments are reported in Sec. 2.4, while concluding remarks are presented in Sec. 2.5.

2.2 Problem Formulation

Our focus in this chapter is on the ultrahigh-dimensional ordinary linear model $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\eta}$, where $\mathbf{y} \in \mathbb{R}^m$, $X \in \mathbb{R}^{m \times n}$ and $n \ggg m$. In the statistics literature,

Algorithm 1: Marginal Correlation-based Screening

Input: $X \in \mathbb{R}^{m \times n}$, $\mathbf{y} \in \mathbb{R}^m$, and $d \in \mathbb{Z}_+$

1: $\mathbf{w} \leftarrow X^\top \mathbf{y}$

2: $\widehat{\mathcal{S}}_d \leftarrow \{i \in [[n]] : |\mathbf{w}_i| \text{ is among the } d \text{ largest of all correlations}\}$

Output: $\widehat{\mathcal{S}}_d \subset [[n]]$ such that $|\widehat{\mathcal{S}}_d| = d$

X is referred to as data/design/observation matrix with the rows of X corresponding to individual observations and the columns of X corresponding to individual features/predictors/variables, $\mathbf{y}$ is referred to as observation/response vector with individual responses given by $\{\mathbf{y}_i\}_{i=1}^m$, $\boldsymbol{\beta}$ is referred to as the parameter vector, and $\boldsymbol{\eta}$ is referred to as modeling error or observation noise. Throughout this chapter, we assume X has unit ℓ_2-norm columns, $\boldsymbol{\beta} \in \mathbb{R}^n$ is k-sparse with $k < m$ (i.e., $\left|\{i \in [[n]] : \boldsymbol{\beta}_i \neq 0\}\right| = k < m$), and $\boldsymbol{\eta} \in \mathbb{R}^n$ is a zero-mean Gaussian vector with (entry-wise) variance σ^2 and covariance $C_{\boldsymbol{\eta}} = \sigma^2 I$. Here, $\boldsymbol{\eta}$ is taken to be Gaussian with covariance $\sigma^2 I$ for the sake of this exposition, but our analysis is trivially generalizable to other noise distributions and/or covariance matrices. Further, we make no a priori assumption on the distribution of X. Finally, we define $\mathcal{S} := \{i \in [[n]] : \boldsymbol{\beta}_i \neq 0\}$ to be the set that indexes the non-zero components of $\boldsymbol{\beta}$. Using this notation, the linear model can equivalently be expressed as

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\eta} = X_{\mathcal{S}}\boldsymbol{\beta}_{\mathcal{S}} + \boldsymbol{\eta}. \tag{2.1}$$

Given (2.1), the goal of variable screening is to reduce the number of variables in the linear model from n (since $n \gg m$) to a moderate scale d ($\ll n$) using a fast and efficient method. Our focus here is on screening methods that satisfy the so-called *sure screening* property [36]; specifically, a method is said to carry out sure screening if the d-dimensional model returned by it is guaranteed with high probability to retain all the columns of X that are indexed by $\mathcal{S}$. In this chapter, we study sure screening using marginal correlations between the response vector and the columns of X. The resulting screening procedure is outlined in Algorithm 1.

The term *sure independence screening* (SIS) was coined in [36] to refer to screening of ultrahigh-dimensional *Gaussian* linear models using Algorithm 1. Our goal in this chapter is to provide an understanding of the screening performance of Algorithm 1 for

arbitrary (and, thus, not just Gaussian) design matrices. We use the term *extended sure independence screening* (ExSIS) to refer to screening of arbitrary linear models using Algorithm 1.

2.3 Sufficient Conditions for Sure Screening

In this section, we derive the most general sufficient conditions for ExSIS of ultrahigh-dimensional linear models. The results reported in this section provide important insights into the workings of ExSIS *without* imposing any statistical priors on X and $\boldsymbol{\beta}$. We begin with a definition of the *screening condition* for the design matrix X.

Definition 2.1 ((k,b)-Screening Condition)**.** Fix an arbitrary $\boldsymbol{\beta} \in \mathbb{R}^n$ that is k-sparse. The (normalized) matrix X satisfies the (k,b)-screening condition if there exists $0 < b(m,n) < \frac{1}{\sqrt{k}}$ such that the following hold:

$$\max_{i \in \mathcal{S}} |\sum_{\substack{j \in \mathcal{S} \\ j \neq i}} X_i^\top X_j \boldsymbol{\beta}_j| \leq b(m,n) \|\boldsymbol{\beta}\|_2, \tag{SC-1}$$

$$\max_{i \in \mathcal{S}^c} |\sum_{j \in \mathcal{S}} X_i^\top X_j \boldsymbol{\beta}_j| \leq b(m,n) \|\boldsymbol{\beta}\|_2. \tag{SC-2}$$

The screening condition is a statement about the collinearity of the independent variables in the design matrix. The parameter $b(m,n)$ in the screening condition captures the similarity between (i) the columns of $X_{\mathcal{S}}$, and (ii) the columns of $X_{\mathcal{S}}$ and $X_{\mathcal{S}^c}$; the smaller the parameter $b(m,n)$ is, the less similar the columns are. Furthermore, since $k < (b(m,n))^{-2}$ in the screening condition, the parameter $b(m,n)$ reflects constraints on the sparsity parameter k.

We now present one of our main screening results for arbitrary design matrices, which highlights the significance of the screening condition and the role of the parameter $b(m,n)$ within ExSIS.

Theorem 2.1 (Sufficient Conditions for ExSIS)**.** *Let* $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\eta}$ *with* $\boldsymbol{\beta}$ *a* k*-sparse vector and the entries of* $\boldsymbol{\eta}$ *independently distributed as* $\mathcal{N}(0, \sigma^2)$*. Define* $\beta_{\min} := \min_{i \in \mathcal{S}} |\boldsymbol{\beta}_i|$ *and* $\tilde{\boldsymbol{\eta}} := X^\top \boldsymbol{\eta}$*, and let* $\mathcal{G}_\eta$ *be the event* $\{\|\tilde{\boldsymbol{\eta}}\|_\infty \leq 2\sqrt{\sigma^2 \log n}\}$*. Suppose* X *satisfies the*

screening condition and assume $\frac{\beta_{\min}}{\|\boldsymbol{\beta}\|_2} > 2b(m,n) + 4\frac{\sqrt{\sigma^2 \log n}}{\|\boldsymbol{\beta}\|_2}$. *Then, conditioned on* $\mathcal{G}_\eta$, *Algorithm 1 satisfies* $\mathcal{S} \subset \widehat{\mathcal{S}}_d$ *as long as* $d \geq \left\lceil \frac{\sqrt{k}}{\frac{\beta_{\min}}{\|\boldsymbol{\beta}\|_2} - 2b(m,n) - \frac{4\sqrt{\sigma^2 \log n}}{\|\boldsymbol{\beta}\|_2}} \right\rceil$.

We refer the reader to Sec. 2.3.2 for a proof of this theorem.

2.3.1 Discussion

Theorem 2.1 highlights the dependence of ExSIS on the observation noise, the ratio $\frac{\beta_{\min}}{\|\boldsymbol{\beta}\|_2}$, the parameter $b(m,n)$, and model sparsity. We first comment on the relationship between ExSIS and observation noise $\boldsymbol{\eta}$. Notice that the statement of Theorem 2.1 is dependent upon the event $\mathcal{G}_\eta$. However, for any $\epsilon > 0$, we have (see, e.g., [50, Lemma 6])

$$\Pr(\|\tilde{\boldsymbol{\eta}}\|_\infty \geq \sigma\epsilon) < \frac{4n}{\epsilon\sqrt{2\pi}} \exp\Big(-\frac{\epsilon^2}{2}\Big). \tag{2.2}$$

Therefore, substituting $\epsilon = 2\sqrt{\log n}$ in (2.2), we obtain

$$\Pr(\mathcal{G}_\eta) \geq 1 - 2(n\sqrt{2\pi \log n})^{-1}. \tag{2.3}$$

Thus, Algorithm 1 possesses the sure screening property in the case of the observation noise $\boldsymbol{\eta}$ distributed as $\mathcal{N}(0, \sigma^2 I)$. We further note from the statement of Theorem 2.1 that the higher the *signal-to-noise ratio* (SNR), defined here as SNR $:= \frac{\|\boldsymbol{\beta}\|_2}{\sigma}$, the more Algorithm 1 can screen irrelevant/inactive variables. It is also worth noting here trivial generalizations of Theorem 2.1 for other noise distributions. In the case of $\boldsymbol{\eta}$ distributed as $\mathcal{N}(0, C_{\boldsymbol{\eta}})$, Theorem 2.1 has σ^2 replaced by the largest eigenvalue of the covariance matrix $C_{\boldsymbol{\eta}}$. In the case of $\boldsymbol{\eta}$ following a non-Gaussian distribution, Theorem 2.1 has $2\sqrt{\sigma^2 \log n}$ replaced by distribution-specific upper bound on $\|X^\top \boldsymbol{\eta}\|_\infty$ that holds with high probability.

In addition to the noise distribution, the performance of ExSIS also seems to be impacted by the *minimum-to-signal ratio* (MSR), defined here as MSR $:= \frac{\beta_{\min}}{\|\boldsymbol{\beta}\|_2} \in \big(0, \frac{1}{\sqrt{k}}\big]$. Specifically, the higher the MSR, the more Algorithm 1 can screen inactive variables. Stated differently, the independent variable with the weakest contribution to the response determines the size of the screened model. Finally, the parameter $b(m,n)$ in the screening condition also plays a central role in characterization of the performance

of ExSIS. First, the smaller the parameter $b(m,n)$, the more Algorithm 1 can screen inactive variables. Second, the smaller the parameter $b(m,n)$, the more independent variables can be active in the original model; indeed, we have from the screening condition that $k < (b(m,n))^{-2}$. Third, the smaller the parameter $b(m,n)$, the lower the smallest allowable value of MSR; indeed, we have from the theorem statement that $\text{MSR} > 2b(m,n) + 4\frac{\sqrt{\sigma^2 \log n}}{\|\boldsymbol{\beta}\|_2}$.

It is evident from the preceding discussion that the screening condition (equivalently, the parameter $b(m,n)$) is one of the most important factors that helps understand the workings of ExSIS and helps quantify its performance. Unfortunately, the usefulness of this knowledge is limited in the sense that the screening condition cannot be utilized in practice. Specifically, the screening condition is defined in terms of the set $\mathcal{S}$, which is of course unknown. We overcome this limitation of Theorem 2.1 in the next chapter by implicitly deriving the screening condition for sub-Gaussian design matrices in Sec. 3.2 and for a class of arbitrary (random or deterministic) design matrices in Sec. 3.3.

2.3.2 Proof of Theorem 2.1

We first provide an outline of the proof of Theorem 2.1, which is followed by its formal proof. Define $p_0 := n$, $\widehat{\mathcal{S}}_{p_0} := [[n]]$, and $t_1 := |\{j \in \widehat{\mathcal{S}}_{p_0} : |\mathbf{w}_j| \geq \min_{i \in \mathcal{S}} |\mathbf{w}_i|\}|$. Next, fix a positive integer $p_1 < p_0$ and define

$$\widehat{\mathcal{S}}_{p_1} := \{i \in \widehat{\mathcal{S}}_{p_0} : |\mathbf{w}_i| \text{ is among the } p_1 \text{ largest of all marginal correlations}\}.$$

The idea is to first derive an initial upper bound on t_1, denoted by $\bar{t}_1$, and then choose $p_1 = \lceil \bar{t}_1 \rceil$; trivially, we have $\mathcal{S} \subset \widehat{\mathcal{S}}_{p_1} \subset \widehat{\mathcal{S}}_{p_0}$. As a result, we get

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\eta} = X_{\widehat{\mathcal{S}}_{p_0}} \boldsymbol{\beta}_{\widehat{\mathcal{S}}_{p_0}} + \boldsymbol{\eta} = X_{\widehat{\mathcal{S}}_{p_1}} \boldsymbol{\beta}_{\widehat{\mathcal{S}}_{p_1}} + \boldsymbol{\eta}. \tag{2.4}$$

Note that while deriving $\bar{t}_1$, we need to ensure $\bar{t}_1 < p_0$; this in turn imposes some conditions on X that also need to be specified. Next, we can repeat the aforementioned steps to obtain $\widehat{\mathcal{S}}_{p_2}$ from $\widehat{\mathcal{S}}_{p_1}$ for a fixed positive integer $p_2 < p_1 < p_0$. Specifically, define

$$\widehat{\mathcal{S}}_{p_2} := \{i \in \widehat{\mathcal{S}}_{p_1} : |\mathbf{w}_i| \text{ is among the } p_2 \text{ largest of all marginal correlations}\}$$

and $t_2 := |\{j \in \widehat{\mathcal{S}}_{p_1} : |\mathbf{w}_j| \geq \min_{i \in \mathcal{S}} |\mathbf{w}_i|\}|$. We can then derive an upper bound on t_2, denoted by $\bar{t}_2$, and then choose $p_2 = \lceil \bar{t}_2 \rceil$; once again, we have $\mathcal{S} \subset \widehat{\mathcal{S}}_{p_2} \subset \widehat{\mathcal{S}}_{p_1} \subset \widehat{\mathcal{S}}_{p_0}$. Notice further that we do require $\bar{t}_2 < p_1$, which again will impose conditions on X.

In similar vein, we can keep on repeating this procedure to obtain a decreasing sequence of numbers $\{\bar{t}_j\}_{j=1}^{i}$ and sets $\widehat{\mathcal{S}}_{p_0} \supset \widehat{\mathcal{S}}_{p_1} \supset \widehat{\mathcal{S}}_{p_2} \supset \ldots \supset \widehat{\mathcal{S}}_{p_i} \supset \mathcal{S}$ as long as $\bar{t}_i < p_{i-1}$, where $\left\{p_j := \lceil \bar{t}_j \rceil\right\}_{j=1}^{i}$ and $i \in \mathbb{Z}_+$. The complete proof of Theorem 2.1 follows from a careful combination of these (analytical) steps. In order for us to be able to do that, however, we need two lemmas. The first lemma provides an upper bound on $t_i = |\{j \in \widehat{\mathcal{S}}_{p_{i-1}} : |\mathbf{w}_j| \geq \min_{i \in \mathcal{S}} |\mathbf{w}_i|\}|$ for $i \in \mathbb{Z}_+$, denoted by $\bar{t}_i$. The second lemma provides conditions on the design matrix X such that $\bar{t}_i < p_{i-1}$. The proof of the theorem follows from repeated application of the two lemmas.

Lemma 2.1. *Fix $i \in \mathbb{Z}_+$ and suppose $\mathcal{S} \subset \widehat{\mathcal{S}}_{p_{i-1}}$, where $|\widehat{\mathcal{S}}_{p_{i-1}}| =: p_{i-1}$ and $p_{i-1} \leq p$. Further, suppose the design matrix X satisfies the (k, b)-screening condition for the k-sparse vector $\boldsymbol{\beta}$ and the event $\mathcal{G}_\eta$ holds true. Finally, define $t_i := |\{j \in \widehat{\mathcal{S}}_{p_{i-1}} : |\mathbf{w}_j| \geq \min_{i \in \mathcal{S}} |\mathbf{w}_i|\}|$. Under these conditions, we have*

$$t_i \leq \frac{p_{i-1} b(m,n) \|\boldsymbol{\beta}\|_2 + \|\boldsymbol{\beta}\|_1 + 2p_{i-1}\sqrt{\sigma^2 \log n}}{\beta_{\min} - b(m,n)\|\boldsymbol{\beta}\|_2 - 2\sqrt{\sigma^2 \log n}} =: \bar{t}_i. \tag{2.5}$$

The proof of this lemma is provided in Appendix 2.6.1. The second lemma, whose proof is given in Appendix 2.6.2, provides conditions on X under which the upper bound derived on t_i for $i \in \mathbb{Z}_+$, denoted by $\bar{t}_i$, is non-trivial.

Lemma 2.2. *Fix $i \in \mathbb{Z}_+$. Suppose $p_{i-1} > \frac{\sqrt{k}}{\frac{\beta_{\min}}{\|\boldsymbol{\beta}\|_2} - 2b(m,n) - \frac{4\sqrt{\sigma^2 \log n}}{\|\boldsymbol{\beta}\|_2}}$ and $\frac{\beta_{\min}}{\|\boldsymbol{\beta}\|_2} > 2b(m,n) + \frac{4\sqrt{\sigma^2 \log n}}{\|\boldsymbol{\beta}\|_2}$. Then, we have $0 < \bar{t}_i < p_{i-1}$.*

We are now ready to present a complete technical proof of Theorem 2.1.

Proof. The idea is to use Lemma 2.1 and Lemma 2.2 *repeatedly* to screen columns of X. Note, however, that this is simply an analytical technique and we do not *actually* need to perform such an iterative procedure to specify d in Algorithm 1. To begin, recall that

we have $p_0 := n$, $\widehat{\mathcal{S}}_{p_0} := [[p_0]]$,

$$\widehat{\mathcal{S}}_{p_1} := \{i \in \widehat{\mathcal{S}}_{p_0} : |\mathbf{w}_i| \text{ is among the } p_1 \text{ largest of all marginal correlations}\},$$

and $p_1 = \lceil \bar{t}_1 \rceil$, where $\bar{t}_1$ is defined in (2.5). By Lemma 2.1 and Lemma 2.2, we have $\mathcal{S} \subset \widehat{\mathcal{S}}_{p_1}$ and $p_1 < p_0$, respectively. Next, given $p_1 > \left\lceil \frac{\sqrt{k}}{\frac{\beta_{\min}}{\|\boldsymbol{\beta}\|_2} - 2b(m,n) - \frac{4\sqrt{\sigma^2 \log n}}{\|\boldsymbol{\beta}\|_2}} \right\rceil$, we can use Lemma 2.1 and Lemma 2.2 to obtain $\widehat{\mathcal{S}}_{p_2}$ from $\widehat{\mathcal{S}}_{p_1}$ in a similar fashion. Specifically, let

$$\widehat{\mathcal{S}}_{p_2} := \{i \in \widehat{\mathcal{S}}_{p_1} : |\mathbf{w}_i| \text{ is among the } p_2 \text{ largest of all marginal correlations}\}$$

and $p_2 = \lceil \bar{t}_2 \rceil$, where $\bar{t}_2$ is defined in (2.5). Then, by Lemma 2.1 and Lemma 2.2, we have $\mathcal{S} \subset \widehat{\mathcal{S}}_{p_2}$ and $p_2 < p_1$, respectively.

Notice that we can keep on repeating this procedure to obtain sub-models $\widehat{\mathcal{S}}_{p_1}, \widehat{\mathcal{S}}_{p_2}, \ldots, \widehat{\mathcal{S}}_{p_l}$ such that $p_l \leq \frac{\sqrt{k}}{\frac{\beta_{\min}}{\|\boldsymbol{\beta}\|_2} - 2b(m,n) - \frac{4\sqrt{\sigma^2 \log p}}{\|\boldsymbol{\beta}\|_2}}$ and $p_{l-1} > \frac{\sqrt{k}}{\frac{\beta_{\min}}{\|\boldsymbol{\beta}\|_2} - 2b(m,n) - \frac{4\sqrt{\sigma^2 \log p}}{\|\boldsymbol{\beta}\|_2}}$. By repeated applications of Lemma 2.1 and Lemma 2.2, we have $\mathcal{S} \subset \widehat{\mathcal{S}}_{p_l}$. Further, we are also guaranteed that $p_l \leq \frac{\sqrt{k}}{\frac{\beta_{\min}}{\|\boldsymbol{\beta}\|_2} - 2b(m,n) - \frac{4\sqrt{\sigma^2 \log n}}{\|\boldsymbol{\beta}\|_2}}$. Thus, we can choose $d \geq \left\lceil \frac{\sqrt{k}}{\frac{\beta_{\min}}{\|\boldsymbol{\beta}\|_2} - 2b(m,n) - \frac{4\sqrt{\sigma^2 \log n}}{\|\boldsymbol{\beta}\|_2}} \right\rceil$ in Algorithm 1 in one shot and have $\mathcal{S} \subset \widehat{\mathcal{S}}_d$. □

2.4 Numerical Experiments

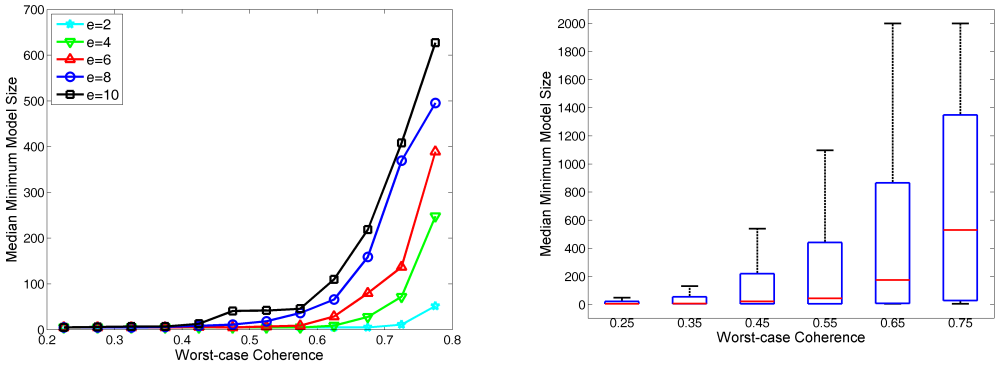
In order to ensure the insights offered by Theorem 2.1 are not mere artifacts of our analysis, we carry out numerical experiments to study the impact of relevant parameters on the screening performance of an *oracle* that has perfect knowledge of the minimum value of d required in Algorithm 1 to ensure $\mathcal{S} \subset \widehat{\mathcal{S}}_d$. In particular, we use these oracle-based experiments to verify the role of $b(m,n)$ and MSR in screening using Algorithm 1, as specified by Theorem 2.1. Before we describe our experiments, let us define the notion of worst-case coherence, μ, of X as defined in [51]: $\mu := \max_{i,j:i \neq j} \left| X_i^\top X_j \right|$. Since worst-case coherence is an indirect measure of pairwise similarity among the columns of X, we use μ as a surrogate for the value of $b(m,n)$ in our experiments.

The design matrix $X \in \mathbb{R}^{m \times n}$ in our experiments is generated such that it consists of independent and identically distributed Gaussian entries, followed by normalization

of the columns of X. Among other parameters, $m = 500$, $n = 2000$, $k = 5$, and $\sigma = 0$ in the experiments. The entries of $\mathcal{S}$ are chosen uniformly at random from $[[n]]$. Furthermore, the non-zero entries in the parameter vector $\boldsymbol{\beta}$ are sampled from a uniform distribution $U[a, e]$; the value of a is set at 1 whereas $e \in [2, 10]$. Finally, the experiments comprise the use of an oracle to find the minimum possible value of d that can be used in Algorithm 1 while ensuring $\mathcal{S} \subset \widehat{\mathcal{S}}_d$. We refer to this minimum value of d as the *minimum model size* (MMS), and we use median of MMS over 400 runs of the experiment as a metric of difficulty of screening.

To analyze the impact of increasing μ (equivalently, $b(m, n)$) and MSR on screening using Algorithm 1, the numerical experiments are repeated for various values of μ and MSR. In particular, the worst-case coherence of X is varied by scaling its largest singular value, followed by normalization of the columns of X, while the MSR is increased by decreasing the value of e. In Fig. 2.1a, we plot the median MMS against μ for different MSR values. The experimental results of the oracle performance offer two interesting insights. First, the median MMS increases with μ; this shows that any analysis for screening using Algorithm 1 needs to account for the similarity between the columns of X. This relationship is captured by the parameter $b(m, n)$ in Theorem 2.1. Second, the difficulty of screening for an oracle increases with decreasing MSR values. This relationship is also reflected in Theorem 2.1: as $\|\boldsymbol{\beta}\|_2$ increases for a fixed e, MSR decreases and the median MMS increases.

More interestingly, if we focus on the plot in Fig. 2.1a for $b = 10$, and we plot the relationship between μ and median MMS along with the interquartile range of MMS for each value of μ, it can be seen that there are instances when the oracle has to select all 2000 predictors to ensure $\mathcal{S} \subset \widehat{\mathcal{S}}_d$ (see boxplot for $\mu = 0.65$ and 0.75). In other words, no screening can be performed at all in these cases. This phenomenon is also reflected in Theorem 2.1: when $b(n, p)$ becomes too large, the condition imposed on MSR is no longer true and our analysis cannot be used for screening using Algorithm 1.



(a) (b)

Figure 2.1: Understanding the limitations of correlation-based screening through the use of an oracle. (a) Relationship between the worst-case coherence and the MMS for various values of MSR. (b) Boxplot of the MMS versus the worst-case coherence for $e = 10$.

2.5 Conclusion

In this chapter, we provided mathematical guarantees for variable screening of arbitrary linear models using a marginal correlation-based approach, without imposing any statistical prior on the linear model. Moreover, our experiments demonstrated that the insights from the main result are reflective of the actual challenges involved with screening of arbitrary linear models using marginal correlations.

2.6 Appendix

2.6.1 Proof of Lemma 2.1

We begin by defining

$$\mathbf{w}^{(i)} := X_{\widehat{\mathcal{S}}_{p_{i-1}}}^{\top} \mathbf{y} = X_{\widehat{\mathcal{S}}_{p_{i-1}}}^{\top} X_{\mathcal{S}} \boldsymbol{\beta}_{\mathcal{S}} + X_{\widehat{\mathcal{S}}_{p_{i-1}}}^{\top} \boldsymbol{\eta} =: \boldsymbol{\xi}^{(i)} + \tilde{\boldsymbol{\eta}}^{(i)} \tag{2.6}$$

where $\mathbf{w}^{(i)} \in \mathbb{R}^{|\widehat{\mathcal{S}}_{p_{i-1}}|}$ measures the correlation of the observation vector y with each column of $X_{\widehat{\mathcal{S}}_{p_{i-1}}}$. To derive an upper bound on t_i, we derive upper and lower bounds on $\sum_{j_1=1}^{p_{i-1}} |\mathbf{w}_{j_1}^{(i)}|$. A simple upper bound on $\sum_{j_1=1}^{p_{i-1}} |\mathbf{w}_{j_1}^{(i)}|$ is:

$$\sum_{j_1=1}^{p_{i-1}} |w_{j_1}^{(i)}| = \sum_{j_1=1}^{p_{i-1}} |\boldsymbol{\xi}_{j_1}^{(i)} + \tilde{\boldsymbol{\eta}}_{j_1}^{(i)}| \leq \sum_{j_1=1}^{p_{i-1}} (|\boldsymbol{\xi}_{j_1}^{(i)}| + |\tilde{\boldsymbol{\eta}}_{j_1}^{(i)}|) = \|\boldsymbol{\xi}^{(i)}\|_1 + \|\tilde{\boldsymbol{\eta}}^{(i)}\|_1. \tag{2.7}$$

Next, we recall that $\mathbf{w} = X^\top y$ and we further define $\boldsymbol{\xi}$ and $\tilde{\boldsymbol{\eta}}$ such that

$$\mathbf{w} = X^\top \mathbf{y} = X^\top X_{\mathcal{S}} \boldsymbol{\beta}_{\mathcal{S}} + X^\top \boldsymbol{\eta} =: \boldsymbol{\xi} + \tilde{\boldsymbol{\eta}}. \tag{2.8}$$

Now define $\mathcal{T}_i := \{j_1 \in \widehat{\mathcal{S}}_{p_{i-1}} : |\mathbf{w}_{j_1}| \geq \min_{j_2 \in \mathcal{S}} |w_{j_2}|\}$. Then, a simple lower bound on $\sum_{j_1=1}^{p_{i-1}} |\mathbf{w}_{j_1}^{(i)}|$ is:

$$\begin{aligned}
\sum_{j_1=1}^{p_{i-1}} |\mathbf{w}_{j_1}^{(i)}| &= \sum_{j_1 \in \widehat{\mathcal{S}}_{p_{i-1}}} |\mathbf{w}_{j_1}| \geq \sum_{j_1 \in \mathcal{T}_i} |\mathbf{w}_{j_1}| \\
&\geq \sum_{j_1 \in \mathcal{T}_i} \min_{j_2 \in \mathcal{T}_i} |\mathbf{w}_{j_2}| \overset{\text{(a)}}{\geq} \sum_{j_1 \in \mathcal{T}_i} \min_{j_2 \in \mathcal{S}} |\mathbf{w}_{j_2}| \\
&= t_i (\min_{j_2 \in \mathcal{S}} |\mathbf{w}_{j_2}|) = t_i (\min_{j_2 \in \mathcal{S}} |\boldsymbol{\xi}_{j_2} + \tilde{\boldsymbol{\eta}}_{j_2}|) \\
&\geq t_i (\min_{j_2 \in \mathcal{S}} |\boldsymbol{\xi}_{j_2}| - \max_{j_2 \in \mathcal{S}} |\tilde{\boldsymbol{\eta}}_{j_2}|),
\end{aligned} \tag{2.9}$$

where (a) follows from definition of $\mathcal{T}_i$. Combining (2.7) with (2.9), we get

$$t_i \leq \frac{\|\boldsymbol{\xi}^{(i)}\|_1 + \|\tilde{\boldsymbol{\eta}}^{(i)}\|_1}{\min_{j_2 \in \mathcal{S}} |\boldsymbol{\xi}_{j_2}| - \max_{j_2 \in \mathcal{S}} |\tilde{\boldsymbol{\eta}}_{j_2}|}. \tag{2.10}$$

We next bound $\|\boldsymbol{\xi}^{(i)}\|_1$, $\max_{j_2 \in \mathcal{S}} |\tilde{\boldsymbol{\eta}}_{j_2}|$, $\|\tilde{\boldsymbol{\eta}}^{(i)}\|_1$ and $\min_{j_2 \in \mathcal{S}} |\boldsymbol{\xi}_{j_2}|$ separately. First, we derive an upper bound on $\|\boldsymbol{\xi}^{(i)}\|_1$:

$$\begin{aligned}
\|\boldsymbol{\xi}^{(i)}\|_1 &= \sum_{j_1 \in \widehat{\mathcal{S}}_{p_{i-1}}} \Big| \sum_{j_2 \in \mathcal{S}} X_{j_1}^\top X_{j_2} \boldsymbol{\beta}_{j_2} \Big| \\
&\overset{\text{(b)}}{=} \sum_{j_1 \in \mathcal{S}} \Big| \sum_{j_2 \in \mathcal{S}} X_{j_1}^\top X_{j_2} \boldsymbol{\beta}_{j_2} \Big| + \sum_{j_1 \in \widehat{\mathcal{S}}_{p_{i-1}} \setminus \mathcal{S}} \Big| \sum_{j_2 \in \mathcal{S}} X_{j_1}^\top X_{j_2} \boldsymbol{\beta}_{j_2} \Big| \\
&\overset{\text{(c)}}{\leq} \sum_{j_1 \in \mathcal{S}} \Big| \sum_{\substack{j_2 \in \mathcal{S} \\ j_2 \neq j_1}} X_{j_1}^\top X_{j_2} \boldsymbol{\beta}_{j_2} \Big| + \sum_{j_1 \in \mathcal{S}} |\boldsymbol{\beta}_{j_1}| + \sum_{j_1 \in \widehat{\mathcal{S}}_{p_{i-1}} \setminus \mathcal{S}} \Big| \sum_{j_2 \in \mathcal{S}} X_{j_1}^\top X_{j_2} \boldsymbol{\beta}_{j_2} \Big| \\
&\leq k \max_{j_1 \in \mathcal{S}} \Big| \sum_{\substack{j_2 \in \mathcal{S} \\ j_2 \neq j_1}} X_{j_1}^\top X_{j_2} \boldsymbol{\beta}_{j_2} \Big| + (p_{i-1} - k) \max_{j_1 \in \widehat{\mathcal{S}}_{p_{i-1}} \setminus \mathcal{S}} \Big| \sum_{j_2 \in \mathcal{S}} X_{j_1}^\top X_{j_2} \boldsymbol{\beta}_{j_2} \Big| + \|\boldsymbol{\beta}\|_1 \\
&\leq k \max_{j_1 \in \mathcal{S}} \Big| \sum_{\substack{j_2 \in \mathcal{S} \\ j_2 \neq j_1}} X_{j_1}^\top X_{j_2} \boldsymbol{\beta}_{j_2} \Big| + (p_{i-1} - k) \max_{j_1 \in \mathcal{S}^c} \Big| \sum_{j_2 \in \mathcal{S}} X_{j_1}^\top X_{j_2} \boldsymbol{\beta}_{j_2} \Big| + \|\boldsymbol{\beta}\|_1,
\end{aligned} \tag{2.11}$$

where (b) follows since $\mathcal{S} \subset \widehat{\mathcal{S}}_{p_{i-1}}$ and (c) follows from the triangle inequality and the fact that the columns of X are unit norm. Next, we have

$$\max_{j_2 \in \mathcal{S}} |\tilde{\boldsymbol{\eta}}_{j_2}| \leq \|\tilde{\boldsymbol{\eta}}\|_\infty \leq 2\sqrt{\sigma^2 \log p}, \tag{2.12}$$

where the last inequality follows from conditioning on $\mathcal{G}_\eta$. Similarly, we have

$$\begin{aligned}\|\tilde{\boldsymbol{\eta}}^{(i)}\|_1 &= \sum_{j_1 \in \widehat{\mathcal{S}}_{p_{i-1}}} |X_{j_1}^\top \boldsymbol{\eta}| \leq \sum_{j_1 \in \widehat{\mathcal{S}}_{p_{i-1}}} \max_{j_2 \in \widehat{\mathcal{S}}_{p_{i-1}}} |X_{j_2}^\top \boldsymbol{\eta}| \\ &= p_{i-1}(\max_{j_2 \in \widehat{\mathcal{S}}_{p_{i-1}}} |X_{j_2}^\top \boldsymbol{\eta}|) \leq 2p_{i-1}\sqrt{\sigma^2 \log p} \end{aligned} \tag{2.13}$$

where the last inequality, again, follows from $\mathcal{G}_\eta$. Last, we lower bound $\min_{j_1 \in \mathcal{S}} |\boldsymbol{\xi}_{j_1}|$ as follows:

$$\begin{aligned}\min_{j_1 \in \mathcal{S}} |\boldsymbol{\xi}_{j_1}| &= \min_{j_1 \in \mathcal{S}} |\sum_{j_2 \in \mathcal{S}} X_{j_1}^\top X_{j_2} \boldsymbol{\beta}_{j_2}| \\ &\overset{\text{(d)}}{=} \min_{j_1 \in \mathcal{S}} |\sum_{\substack{j_2 \in \mathcal{S} \\ j_2 \neq j_1}} X_{j_1}^\top X_{j_2} \boldsymbol{\beta}_{j_2} + \boldsymbol{\beta}_{j_1}| \\ &\geq \min_{j_1 \in \mathcal{S}} |\boldsymbol{\beta}_{j_1}| - \max_{j_1 \in \mathcal{S}} |\sum_{\substack{j_2 \in \mathcal{S} \\ j_2 \neq j_1}} X_{j_1}^\top X_{j_2} \boldsymbol{\beta}_{j_2}| = \beta_{\min} - \max_{j_1 \in \mathcal{S}} |\sum_{\substack{j_2 \in \mathcal{S} \\ j_2 \neq j_1}} X_{j_1}^\top X_{j_2} \boldsymbol{\beta}_{j_2}| \end{aligned} \tag{2.14}$$

where (d) follows because the columns of X are unit norm. Combining (2.11), (2.12), (2.13), (2.14) with (2.10), we obtain

$$t_i \leq \frac{k \max_{j_1 \in \mathcal{S}} |\sum_{\substack{j_2 \in \mathcal{S} \\ j_2 \neq j_1}} X_{j_1}^\top X_{j_2} \boldsymbol{\beta}_{j_2}| + (p_{i-1} - k) \max_{j_1 \in \mathcal{S}^c} |\sum_{j_2 \in \mathcal{S}} X_{j_1}^\top X_{j_2} \boldsymbol{\beta}_{j_2}| + \|\boldsymbol{\beta}\|_1 + 2p_{i-1}\sqrt{\sigma^2 \log p}}{\beta_{\min} - \max_{j_1 \in \mathcal{S}} |\sum_{\substack{j_1 \in \mathcal{S} \\ j_2 \neq j_1}} X_{j_1}^\top X_{j_2} \boldsymbol{\beta}_{j_2}| - 2\sqrt{\sigma^2 \log p}}. \tag{2.15}$$

Assuming the (k, b)-screening condition for the matrix X holds, we finally obtain

$$\begin{aligned} t_i &\leq \frac{kb(n,p) + (p_{i-1} - k)b(n,p) + \|\boldsymbol{\beta}\|_1 + 2p_{i-1}\sqrt{\sigma^2 \log p}}{\beta_{\min} - b(n,p) - 2\sqrt{\sigma^2 \log p}} \\ &= \frac{p_{i-1}b(n,p) + \|\boldsymbol{\beta}\|_1 + 2p_{i-1}\sqrt{\sigma^2 \log p}}{\beta_{\min} - b(n,p) - 2\sqrt{\sigma^2 \log p}}. \end{aligned} \tag{2.16}$$

This completes the proof of the lemma. □

2.6.2 Proof of Lemma 2.2

For $\bar{t}_i < p_{i-1}$, we need

$$\begin{aligned} &\frac{p_{i-1}b(n,p)\|\boldsymbol{\beta}\|_2 + \|\boldsymbol{\beta}\|_1 + 2p_{i-1}\sqrt{\sigma^2 \log p}}{\beta_{\min} - b(n,p)\|\boldsymbol{\beta}\|_2 - 2\sqrt{\sigma^2 \log p}} < p_{i-1} \\ \Leftrightarrow\ & p_{i-1} > \frac{\frac{\|\boldsymbol{\beta}\|_1}{\|\boldsymbol{\beta}\|_2}}{\frac{\beta_{\min}}{\|\boldsymbol{\beta}\|_2} - 2b(n,p) - \frac{4\sqrt{\sigma^2 \log p}}{\|\boldsymbol{\beta}\|_2}}. \end{aligned} \tag{2.17}$$

Since $\|\boldsymbol{\beta}\|_1 \leq \sqrt{k}\|\boldsymbol{\beta}\|_2$, we have

$$p_{i-1} > \frac{\sqrt{k}}{\frac{\beta_{\min}}{\|\boldsymbol{\beta}\|_2} - 2b(n,p) - \frac{4\sqrt{\sigma^2 \log p}}{\|\boldsymbol{\beta}\|_2}} \tag{2.18}$$

as a sufficient condition for (2.17). Thus, (2.18) is a sufficient condition for $\bar{t}_i < p_{i-1}$. $\square$

Chapter 3

Variable Screening in Ultrahigh-dimensional Random and Arbitrary Linear Models

In the previous chapter, we showed that a condition—termed the screening condition—is sufficient for successful correlation-based screening of linear models. In this chapter, we explicitly derive the screening condition for two families of linear models, namely, sub-Gaussian linear models and arbitrary (random or deterministic) linear models. In the process, we establish that—under appropriate conditions—it is possible to reduce the dimension of an ultrahigh-dimensional, arbitrary linear model to almost the sample size even when the number of active variables scales almost linearly with the sample size.

3.1 Introduction and Our Contributions

The previous chapter provided a simple, distribution-agnostic sufficient condition—termed the *screening condition*—for (marginal correlation-based) screening of linear models. This sufficient condition, which succinctly captures joint interactions among both the active and the inactive variables, was then leveraged to explicitly characterize the performance of ExSIS as a function of various problem parameters, including noise variance, the ratio $\frac{\beta_{\min}}{\|\boldsymbol{\beta}\|_2}$, and model sparsity. Despite the theoretical usefulness of the screening condition, it cannot be explicitly verified in polynomial time for any given linear model. This is reminiscent of related conditions such as the *incoherence condition* [52], the *irrepresentable condition* [53], the *restricted isometry property* [54], and the *restricted eigenvalue condition* [55] studied in the literature on high-dimensional linear models.

In order to overcome this limitation of the screening condition, we explicitly derive

it for two families of linear models. The first family corresponds to sub-Gaussian linear models, in which the independent variables are independently drawn from (possibly different) sub-Gaussian distributions. We show that the ExSIS results for this family of linear models generalize the SIS results derived in [36] for normally distributed linear models. The second family corresponds to arbitrary (random or deterministic) linear models in which the (empirical) correlations between independent variables satisfy certain polynomial-time verifiable conditions. The ExSIS results for this family of linear models establish that, under appropriate conditions, it is possible to reduce the dimension of an ultrahigh-dimensional linear model to almost the sample size even when the number of active variables scales almost linearly with the sample size. This, to the best of our knowledge, is the first screening result that provides such explicit and optimistic guarantees *without* imposing a statistical prior on the distribution of the independent variables.

The rest of this chapter is organized as follows. In Sec. 3.2, we derive the screening condition for sub-Gaussian linear models and discuss the resulting ExSIS guarantees in relation to prior work. In Sec. 3.3, we derive the screening condition for arbitrary linear models based on the correlations between independent variables and discuss implications of the derived ExSIS results. Finally, results of extensive numerical experiments on both synthetic and real data are reported in Sec. 3.4, while concluding remarks are presented in Sec. 3.5.

3.2 Screening of Sub-Gaussian Design Matrices

In this section, we characterize the implications of Theorem 2.1 for ExSIS of the family of sub-Gaussian design matrices. As noted in Sec. 2.3, this effort primarily involves establishing the screening condition for sub-Gaussian matrices and specifying the parameter $b(m, n)$ for such matrices. We begin by first recalling the definition of a sub-Gaussian random variable.

Definition 3.1. A zero-mean random variable $\mathcal{X}$ is said to follow a sub-Gaussian distribution $subG(b_0)$ if there exists a sub-Gaussian parameter $b_0 > 0$ such that $\mathbb{E}[\exp(\lambda\mathcal{X})] \leq$

$\exp\left(\frac{b_0^2\lambda^2}{2}\right)$ for all $\lambda \in \mathbb{R}$.

In words, a $subG(b_0)$ random variable is one whose moment generating function is dominated by that of a $\mathcal{N}(0, b_0^2)$ random variable. Some common examples of sub-Gaussian random variables include:

- $\mathcal{X} \sim \mathcal{N}(0, b_0^2) \;\Rightarrow\; \mathcal{X} \sim subG(b_0)$.
- $\mathcal{X} \sim \text{unif}(-b_0, b_0) \;\Rightarrow\; \mathcal{X} \sim subG(b_0)$.
- $|\mathcal{X}| \leq b_0, \mathbb{E}[\mathcal{X}] = 0 \;\Rightarrow\; \mathcal{X} \sim subG(b_0)$.
- $\mathcal{X} \sim \begin{cases} b_0, & \text{with prob. } \frac{1}{2}, \\ -b_0, & \text{with prob. } \frac{1}{2}, \end{cases} \Rightarrow\; \mathcal{X} \sim subG(b_0)$.

Our focus in this section is on design matrices in which entries are first independently drawn from sub-Gaussian distributions and then the columns are normalized. In contrast to prior works, however, we do not require the (pre-normalized) entries to be identically distributed. Rather, we allow each independent variable to be distributed as a sub-Gaussian random variable with a different sub-Gaussian parameter. Thus, the ExSIS analysis of this section is applicable to design matrices in which different columns might have different sub-Gaussian distributions. It is also straightforward to extend our analysis to the case where all (and not just across column) entries of the design matrix are non-identically distributed; we do not focus on this extension in here for the sake of notational clarity.

3.2.1 Main Result

The ExSIS of linear models involving sub-Gaussian design matrices mainly requires establishing the screening condition and characterization of the parameter $b(m, n)$ for sub-Gaussian matrices. We accomplish this by individually deriving (SC-1) and (SC-2) in Definition 2.1 for sub-Gaussian design matrices in the following two lemmas.

Lemma 3.1. *Let $V = [V_{i,j}]$ be an $m \times n$ matrix with the entries $\{V_{i,j}\}_{i,j=1}^{m,n}$ independently distributed as $subG(b_j)$ with variances $\mathbb{E}[V_{i,j}^2] = \sigma_j^2$. Suppose the design matrix X*

is obtained by normalizing the columns of V, i.e., $X = V\,diag(1/\|V_1\|_2, \ldots, 1/\|V_n\|_2)$. Finally, fix an arbitrary $\boldsymbol{\beta} \in \mathbb{R}^n$ that is k-sparse, define $\frac{\sigma_*}{b_*} := \min_{j \in \mathcal{S}} \frac{\sigma_j}{b_j}$, and let $\log n \leq \frac{m}{16}(\frac{\sigma_*}{4b_*})^4$. Then, with probability exceeding $1 - \frac{2k^2}{n^2}$, we have

$$\max_{i \in \mathcal{S}} |\sum_{\substack{j \in \mathcal{S} \\ j \neq i}} X_i^\top X_j \boldsymbol{\beta}_j| \leq \sqrt{\frac{8 \log n}{m}} \Big(\frac{b_*}{\sigma_*} \Big) \|\boldsymbol{\beta}\|_2.$$

Lemma 3.2. *Let $V = [V_{i,j}]$ be an $m \times n$ matrix with the entries $\{V_{i,j}\}_{i,j=1}^{m,n}$ independently distributed as subG(b_j) with variances $\mathbb{E}[V_{i,j}^2] = \sigma_j^2$. Suppose the design matrix X is obtained by normalizing the columns of V, i.e., $X = V\,diag(1/\|V_1\|_2, \ldots, 1/\|V_n\|_2)$. Finally, fix an arbitrary $\boldsymbol{\beta} \in \mathbb{R}^n$ that is k-sparse, define $\frac{\sigma_*}{b_*} := \min_{j \in \mathcal{S}} \frac{\sigma_j}{b_j}$, and let $\log n \leq \frac{m}{16}(\frac{\sigma_*}{4b_*})^4$. Then, with probability exceeding $1 - \frac{2(k+1)(n-k)}{n^2}$, we have*

$$\max_{i \in \mathcal{S}^c} |\sum_{j \in \mathcal{S}} X_i^\top X_j \boldsymbol{\beta}_j| \leq \sqrt{\frac{8 \log n}{m}} \Big(\frac{b_*}{\sigma_*} \Big) \|\boldsymbol{\beta}\|_2.$$

The proofs of Lemma 3.1 and Lemma 3.2 are provided in Appendix 3.6.1 and Appendix 3.6.2, respectively. It now follows from a simple union bound argument that the screening condition holds for sub-Gaussian design matrices with probability exceeding $1 - 2(k+1)n^{-1}$. In particular, we have from Lemma 3.1 and Lemma 3.2 that $b(m,n) = \sqrt{\frac{8 \log n}{m}}(\frac{b_*}{\sigma_*})$ for sub-Gaussian matrices. We can now use this knowledge and Theorem 2.1 to provide the main result for ExSIS of ultrahigh-dimensional linear models involving sub-Gaussian design matrices.

Theorem 3.1 (ExSIS and Sub-Gaussian Matrices)**.** *Let $V = [V_{i,j}]$ be an $m \times n$ matrix with the entries $\{V_{i,j}\}_{i,j=1}^{m,n}$ independently distributed as subG(b_j) with variances $\mathbb{E}[V_{i,j}^2] = \sigma_j^2$. Suppose the design matrix X is obtained by normalizing the columns of V, i.e., $X = V\,diag(1/\|V_1\|_2, \ldots, 1/\|V_n\|_2)$. Next, let $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\eta}$ with $\boldsymbol{\beta}$ a k-sparse vector and the entries of $\boldsymbol{\eta}$ independently distributed as $\mathcal{N}(0, \sigma^2)$. Finally, define $\frac{\sigma_*}{b_*} := \min_{j \in \mathcal{S}} \frac{\sigma_j}{b_j}$ and $\beta_{\min} := \min_{i \in \mathcal{S}} |\boldsymbol{\beta}_i|$, and let $\log n \leq \frac{m}{16}(\frac{\sigma_*}{4b_*})^4$ and $\frac{\beta_{\min}}{\|\boldsymbol{\beta}\|_2} > 2\sqrt{\frac{8 \log n}{m}}(\frac{b_*}{\sigma_*}) + 4\frac{\sqrt{\sigma^2 \log n}}{\|\boldsymbol{\beta}\|_2}$. Then*

Algorithm 1 guarantees $\mathcal{S} \subset \widehat{\mathcal{S}}_d$ *with probability exceeding* $1 - 2(k+2)n^{-1}$ *as long as*

$$d \geq \left\lceil \frac{\sqrt{k}}{\frac{\beta_{\min}}{\|\boldsymbol{\beta}\|_2} - 2\sqrt{\frac{8\log n}{m}}\left(\frac{b_*}{\sigma_*}\right) - \frac{4\sqrt{\sigma^2 \log n}}{\|\boldsymbol{\beta}\|_2}} \right\rceil.$$

Proof. Let $\mathcal{G}_p$ be the event that the design matrix X satisfies the screening condition with parameter $b(m,n) = \sqrt{\frac{8\log n}{m}}(\frac{b_*}{\sigma_*})$. Further, let $\mathcal{G}_\eta$ be the event as defined in Theorem 2.1. It then follows from Lemma 3.1, Lemma 3.2, (2.3), and the union bound that the event $\mathcal{G}_p \cap \mathcal{G}_\eta$ holds with probability exceeding $1-2(k+2)n^{-1}$. The advertised claim now follows directly from Theorem 2.1. □

3.2.2 Discussion

Since Theorem 3.1 follows from Theorem 2.1, it shares many of the insights discussed in Sec. 2.3.1. In particular, Theorem 3.1 allows for exponential scaling of the number of independent variables, $\log n \leq \frac{m}{16}(\frac{\sigma_*}{4b_*})^4$, and dictates that the number of independent variables, d, retained after the screening stage be increased with an increase in the sparsity level and/or the number of independent variables, while it can be decreased with an increase in the SNR, MSR, and/or the number of samples. Notice that the lower bound on d in Theorem 3.1 does require knowledge of the sparsity level. However, this limitation can be overcome in a straightforward manner, as shown below.

Corollary 1. *Let* $V = [V_{i,j}]$ *be an* $m \times n$ *matrix with the entries* $\{V_{i,j}\}_{i,j=1}^{m,n}$ *independently distributed as subG*(b_j^2) *with variances* $\mathbb{E}[V_{i,j}^2] = \sigma_j^2$. *Suppose the design matrix* X *is obtained by normalizing the columns of* V, *i.e.,* $X = V\,diag(1/\|V_1\|_2, \ldots, 1/\|V_n\|_2)$. *Next, let* $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\eta}$ *with* $\boldsymbol{\beta}$ *a* k*-sparse vector and the entries of* $\boldsymbol{\eta}$ *independently distributed as* $\mathcal{N}(0, \sigma^2)$. *Further, define* $\frac{\sigma_*}{b_*} := \min_{j \in \mathcal{S}} \frac{\sigma_j}{b_j}$ *and* $\beta_{\min} := \min_{i \in \mathcal{S}} |\boldsymbol{\beta}_i|$. *Finally, let* $k \leq \frac{m}{\log n}$, $\log n \leq \frac{m}{16}(\frac{\sigma_*}{4b_*})^4$, *and* $\frac{\beta_{\min}}{\|\boldsymbol{\beta}\|_2} > 2c_1\sqrt{\frac{8\log n}{m}}(\frac{b_*}{\sigma_*}) + 4c_2\frac{\sqrt{\sigma^2 \log n}}{\|\boldsymbol{\beta}\|_2}$ *for some constants* $c_1, c_2 > 2$. *Then Algorithm 1 guarantees* $\mathcal{S} \subset \widehat{\mathcal{S}}_d$ *with probability exceeding* $1 - 2(k+2)n^{-1}$ *as long as* $d \geq \left\lceil \frac{m}{\log n} \right\rceil$.

Proof. Theorem 3.1 and the condition $\frac{\beta_{\min}}{\|\boldsymbol{\beta}\|_2} > 2c_1\sqrt{\frac{8\log n}{m}}\left(\frac{b_*}{\sigma_*}\right) + \frac{4c_2\sqrt{\sigma^2\log n}}{\|\boldsymbol{\beta}\|_2}$ dictates

$$d \geq \left\lceil \frac{\sqrt{k}}{2(c_1-1)\sqrt{\frac{8\log n}{m}}\left(\frac{b_*}{\sigma_*}\right) + \frac{4(c_2-1)\sqrt{\sigma^2\log n}}{\|\boldsymbol{\beta}\|_2}} \right\rceil \tag{3.1}$$

for sure screening of sub-Gaussian design matrices. The claim now follows by noting that $d \geq \left\lceil \frac{m}{\log n} \right\rceil$ is a sufficient condition for (3.1) since $k \leq \frac{m}{\log n}$ and $\sigma_* \leq b_*$ for sub-Gaussian random variables. □

A few remarks are in order now concerning our analysis of ExSIS for sub-Gaussian design matrices and that of SIS for random matrices in the existing literature. To this end, we focus on the results reported in [36], which is one of the most influential SIS works. In contrast to the screening condition presented in the previous chapter, the analysis in [36] is carried out for design matrices that satisfy a certain concentration property. Since the said concentration property has only been shown in [36] to hold for Gaussian matrices, our discussion in the following is limited to Gaussian design matrices with independent entries.

The SIS results reported in [36] hold under four specific conditions. In particular, Condition 3 in [36] requires that: (i) the variance of the response variable is $\mathcal{O}(1)$, (ii) $\beta_{\min} \geq \frac{c_\kappa}{m^\kappa}$ for some $c_\kappa > 0$, $\kappa \geq 0$, and (iii) $\min_{i\in\mathcal{S}} |\mathrm{cov}(\boldsymbol{\beta}_i^{-1}Y, X_i)| \geq c_3$ for some $c_3 > 0$. Notice, however, that the $\mathcal{O}(1)$ variance condition is equivalent to having $\|\boldsymbol{\beta}\|_2 = \mathcal{O}(1)$. Our analysis, in contrast, imposes no such restriction. Rather, Theorem 3.1 shows that marginal correlation-based sure screening is fundamentally affected by the MSR $\frac{\beta_{\min}}{\|\boldsymbol{\beta}\|_2}$. While Theorem 3.1 is only concerned with sufficient conditions, numerical experiments reported in Sec. 3.4 confirm this dependence. Next, notice that $\max_{i\in\mathcal{S}} |\sum_{\substack{j\in\mathcal{S}\\ j\neq i}} \mathrm{cov}(X_i, X_j)\boldsymbol{\beta}_j| \leq 1 - c_3$ implies $\min_{i\in\mathcal{S}} |\mathrm{cov}(\boldsymbol{\beta}_i^{-1}Y, X_i)| \geq c_3$. It therefore follows that (SC-1) in the screening condition is a non-statistical variant of the condition $\min_{i\in\mathcal{S}} |\mathrm{cov}(\boldsymbol{\beta}_i^{-1}Y, X_i)| \geq c_3$ in [36].

We next assume $\sigma = 0$ for the sake of simplicity of argument and explicitly compare Theorem 3.1 and [36, Theorem 1] for the case of Gaussian design matrices with independent entries. Similar to [36], we also impose the condition $\|\boldsymbol{\beta}\|_2 = \mathcal{O}(1)$ for

comparison purposes. In this setting, both the theorems guarantee sure screening with high probability. In [36, Theorem 1], this requires $\beta_{\min} = \Omega(n^{-\kappa})$ for $\kappa < 1/2$ and $\log n = \mathcal{O}(m^{\alpha})$ for some $\alpha \in (0, 1-2\kappa)$. It is, however, easy to verify that substituting $\beta_{\min} = \Omega(m^{-\kappa})$ and $\log n = \mathcal{O}(m^{\alpha})$ in Theorem 3.1 results in identical constraints of $\kappa < 1/2$ and $\alpha < 1-2\kappa$ for our analysis. Next, [36, Theorem 1] also imposes the sparsity constraint $k = \mathcal{O}(m^{2\kappa})$ for the sure screening result to hold. However, the condition $\log n = \mathcal{O}(m^{\alpha})$ with $\alpha \in (0, 1-2\kappa)$ reduces this constraint to $k = \mathcal{O}(n^{1-\alpha}) = \mathcal{O}\left(\frac{m}{\log n}\right)$, which matches the sparsity constraint imposed by Theorem 3.1 (cf. Corollary 1). To summarize, the ExSIS results derived in this chapter coincide with the ones in [36] for the case of Gaussian design matrices. However, our results are more general in the sense that they explicitly bring out the dependence of Algorithm 1 on the SNR and the MSR, which is something missing in [36], and they are applicable to sub-Gaussian design matrices.

3.3 Screening of Arbitrary Design Matrices

The ExSIS analysis in Sec. 3.2 specializes Theorem 2.1 for sub-Gaussian design matrices. But what about the design matrices in which either the entries do not follow sub-Gaussian distributions or the statistical distributions of entries are unknown? We address this particular question in this section by deriving verifiable sufficient conditions that guarantee the screening condition for any arbitrary (random or deterministic) design matrix. These sufficient conditions are presented in terms of two measures of similarity among the columns of a design matrix. These measures, termed *worst-case coherence* and *average coherence*, are defined as follows.

Definition 3.2 (Worst-case and Average Coherences)**.** Let X be an $m \times n$ matrix with unit ℓ_2-norm columns. The worst-case coherence of X is denoted by μ and is defined as [51]:

$$\mu := \max_{i,j:i \neq j} \left| X_i^{\top} X_j \right|.$$

On the other hand, the average coherence of X is denoted by ν and is defined as [50]:

$$\nu := \frac{1}{p-1}\max_i \left| \sum_{j:j\neq i} X_i^\top X_j \right|.$$

Notice that both the worst-case and the average coherences are readily computable in polynomial time. Heuristically, the worst-case coherence is an indirect measure of pairwise similarity among the columns of X: $\mu \in [0,1]$ with $\mu \searrow 0$ as the columns of X become less similar and $\mu \nearrow 1$ as at least two columns of X become more similar. The average coherence, on the other hand, is an indirect measure of both the collective similarity among the columns of X and the spread of the columns of X within the unit sphere: $\nu \in [0,\mu]$ with $\nu \searrow 0$ as the columns of X become more spread out in $\mathbb{R}^m$ and $\nu \nearrow \mu$ as the columns of X become less spread out. We refer the reader to [56] for further discussion of these two measures as well as their values for commonly encountered matrices.

We are now ready to describe the main results of this section. The first result connects the screening condition to the worst-case coherence. We will see, however, that this result suffers from the so-called square-root bottleneck: ExSIS analysis based solely on the worst-case coherence can, at best, handle $k = \mathcal{O}(\sqrt{m})$ scaling of the sparsity parameter. The second result overcomes this bottleneck by connecting the screening condition to both worst-case and average coherences. The caveat here is that this result imposes a mild statistical prior on the set $\mathcal{S}$.

3.3.1 ExSIS and the Worst-case Coherence

We begin by relating the worst-case coherence of an arbitrary design matrix X with unit-norm columns to the screening condition.

Lemma 3.3 (Worst-case Coherence and the Screening Condition)**.** *Let X be an $m \times n$ design matrix with unit-norm columns. Then, we have*

$$\max_{i\in\mathcal{S}} |\sum_{\substack{j\in\mathcal{S}\\ j\neq i}} X_i^\top X_j \boldsymbol{\beta}_j| \leq \mu\sqrt{k}\|\boldsymbol{\beta}\|_2, \text{ and}$$

$$\max_{i\in\mathcal{S}^c} |\sum_{j\in\mathcal{S}} X_i^\top X_j \boldsymbol{\beta}_j| \leq \mu\sqrt{k}\|\boldsymbol{\beta}\|_2.$$

The proof of this lemma is provided in Appendix 3.6.3. It follows from Lemma 3.3 that a design matrix satisfies the screening condition with parameter $b(m,n) = \mu\sqrt{k}$ as long as $\mu < k^{-1} \Leftrightarrow k < \mu^{-1}$. We now combine this implication of Lemma 3.3 with Theorem 2.1 to provide a result for ExSIS of arbitrary linear models.

Theorem 3.2. *Let $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\eta}$ with $\boldsymbol{\beta}$ a k-sparse vector and the entries of $\boldsymbol{\eta}$ independently distributed as $\mathcal{N}(0,\sigma^2)$. Suppose $k < \mu^{-1}$ and $\frac{\beta_{\min}}{\|\boldsymbol{\beta}\|_2} > 2\mu\sqrt{k} + 4\frac{\sqrt{\sigma^2 \log p}}{\|\boldsymbol{\beta}\|_2}$. Then, Algorithm 1 satisfies $\mathcal{S} \subset \widehat{\mathcal{S}}_d$ with probability exceeding $1 - 2(n\sqrt{2\pi \log n})^{-1}$ as long as $d \geq \left\lceil \frac{\sqrt{k}}{\frac{\beta_{\min}}{\|\boldsymbol{\beta}\|_2} - 2\mu\sqrt{k} - \frac{4\sqrt{\sigma^2 \log n}}{\|\boldsymbol{\beta}\|_2}} \right\rceil$.*

The proof of this theorem follows directly from Lemma 3.3 and Theorem 2.1. Next, a straightforward corollary of Theorem 3.2 shows that ExSIS of arbitrary linear models can in fact be carried out without explicit knowledge of the sparsity parameter k.

Corollary 2. *Let $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\eta}$ with $\boldsymbol{\beta}$ a k-sparse vector and the entries of $\boldsymbol{\eta}$ independently distributed as $\mathcal{N}(0,\sigma^2)$. Suppose $n \geq 2m$, $k < \mu^{-1}$, and $\frac{\beta_{\min}}{\|\boldsymbol{\beta}\|_2} > 2c_1\mu\sqrt{k} + 4c_2\frac{\sqrt{\sigma^2 \log n}}{\|\boldsymbol{\beta}\|_2}$ for some $c_1, c_2 > 2$. Then, Algorithm 1 satisfies $\mathcal{S} \subset \widehat{\mathcal{S}}_d$ with probability exceeding $1 - 2(n\sqrt{2\pi \log n})^{-1}$ as long as $d \geq \lceil \sqrt{m} \rceil$.*

Proof. Under the assumption of $\frac{\beta_{\min}}{\|\boldsymbol{\beta}\|_2} > 2c_1\mu\sqrt{k} + \frac{4c_2\sqrt{\sigma^2 \log n}}{\|\boldsymbol{\beta}\|_2}$, notice that

$$d \geq \left\lceil \frac{\sqrt{k}}{2(c_1 - 1)\mu\sqrt{k} + \frac{4(c_2 - 1)\sqrt{\sigma^2 \log n}}{\|\boldsymbol{\beta}\|_2}} \right\rceil \tag{3.2}$$

is a sufficient condition for $d \geq \left\lceil \frac{\sqrt{k}}{\frac{\beta_{\min}}{\|\boldsymbol{\beta}\|_2} - 2\mu\sqrt{k} - \frac{4\sqrt{\sigma^2 \log n}}{\|\boldsymbol{\beta}\|_2}} \right\rceil$. Further, note that $d \geq \lceil (2\mu)^{-1} \rceil$ is a sufficient condition for (3.2). Next, since $n \geq 2m$, we also have $\mu^{-1} \leq \sqrt{2m}$ from the Welch bound on the worst-case coherence of design matrices [57]. Thus, $d \geq \lceil \sqrt{m} \rceil$ is a sufficient condition for $d \geq \left\lceil \frac{\sqrt{k}}{\frac{\beta_{\min}}{\|\boldsymbol{\beta}\|_2} - 2\mu\sqrt{k} - \frac{4\sqrt{\sigma^2 \log n}}{\|\boldsymbol{\beta}\|_2}} \right\rceil$. □

It is interesting to compare this result for arbitrary linear models with Corollary 1 for sub-Gaussian linear models. Corollary 1 requires the size of the screened model to scale as $\mathcal{O}(m/\log n)$, whereas this result requires d to scale only as $\mathcal{O}(\sqrt{m})$. While

this may seem to suggest that Corollary 2 is better than Corollary 1, such an observation ignores the respective constraints on the sparsity parameter k in the two results. Specifically, Corollary 1 allows for almost linear scaling of the sparsity parameter, $k = \mathcal{O}(m/\log n)$, whereas Corollary 2 suffers from the so-called square-root bottleneck: $k = \mathcal{O}(\mu^{-1}) \Rightarrow k = \mathcal{O}(\sqrt{m})$ because of the Welch bound. Stated differently, Corollary 2 fails to specialize to Corollary 1 for the case of X being a sub-Gaussian design matrix. We overcome this limitation of the results of this section by adding the average coherence into the mix and imposing a statistical prior on the true model $\mathcal{S}$ in the next section.

3.3.2 ExSIS and the Coherence Property

In order to break the square-root bottleneck for ExSIS of arbitrary linear models, we first define the notion of the coherence property.

Definition 3.3 (The Coherence Property)**.** An $m \times n$ design matrix X with unit-norm columns is said to obey the coherence property if there exists a constant $c_\mu > 0$ such that $\mu < \frac{1}{c_\mu \sqrt{\log n}}$ and $\nu < \frac{\mu}{\sqrt{m}}$.

Heuristically, the coherence property, which was first introduced in [50], requires the independent variables to be sufficiently (marginally and jointly) uncorrelated. Notice that, unlike many conditions in high-dimensional statistics (see, e.g., [52–55]), the coherence property is explicitly certifiable in polynomial time for any given design matrix. We now establish that the coherence property implies the design matrix satisfies the screening condition with high probability, where the probability is with respect to uniform prior on the true model $\mathcal{S}$.

Lemma 3.4 (Coherence Property and the Screening Condition)**.** *Let X be an $m \times n$ design matrix that satisfies the coherence property with $c_\mu > 10\sqrt{2}$, and suppose $n \geq \max\{2m, \exp(5)\}$ and $k \leq \frac{m}{\log n}$. Further, assume $\mathcal{S}$ is drawn uniformly at random from*

k-subsets of $[[n]]$. *Then, with probability exceeding* $1 - 4n^{-1}$, *we have*

$$\max_{i \in \mathcal{S}} |\sum_{\substack{j \in \mathcal{S} \\ j \neq i}} X_i^\top X_j \boldsymbol{\beta}_j| \leq c_\mu \mu \sqrt{\log n} \|\boldsymbol{\beta}\|_2, \quad \text{and}$$

$$\max_{i \in \mathcal{S}^c} |\sum_{j \in \mathcal{S}} X_i^\top X_j \boldsymbol{\beta}_j| \leq c_\mu \mu \sqrt{\log n} \|\boldsymbol{\beta}\|_2.$$

The proof of this lemma is provided in Appendix 3.6.4. Lemma 3.4 implies that a design matrix that obeys the coherence property also satisfies the screening condition for *most* models with $b(m,n) = c_\mu \mu \sqrt{\log n}$ as long as $\mu < c_\mu^{-1}(k \log n)^{-1/2} \Leftrightarrow k < \frac{\mu^{-2}}{c_\mu^2 \log n}$. Comparing this with Lemma 3.3 and the resulting constraint $k < \mu^{-1}$ for the screening condition to hold in the case of arbitrary design matrices, we see that—at the expense of uniform prior on the true model—the coherence property results in a better bound on the screening parameter as long as $\log n = \mathcal{O}(\mu^{-1})$. We can now utilize Lemma 3.4 along with Theorem 2.1 to provide an improved result for ExSIS of arbitrary linear models.

Theorem 3.3. *Let* $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\eta}$ *with* $\boldsymbol{\beta}$ *a* k*-sparse vector and the entries of* $\boldsymbol{\eta}$ *independently distributed as* $\mathcal{N}(0, \sigma^2)$. *Further, assume* X *satisfies the coherence property with* $c_\mu > 10\sqrt{2}$ *and* $\mathcal{S}$ *is drawn uniformly at random from* k*-subsets of* $[[n]]$. *Finally, suppose* $n \geq \max\{2m, \exp(5)\}$, $k < \frac{\mu^{-2}}{c_\mu^2 \log n}$, *and* $\frac{\beta_{\min}}{\|\boldsymbol{\beta}\|_2} > 2c_\mu \mu \sqrt{\log n} + \frac{4\sqrt{\sigma^2 \log n}}{\|\boldsymbol{\beta}\|_2}$. *Then, Algorithm 1 satisfies* $\mathcal{S} \subset \widehat{\mathcal{S}}_d$ *with probability exceeding* $1 - 6n^{-1}$ *as long as* $d \geq \left\lceil \frac{\sqrt{k}}{\frac{\beta_{\min}}{\|\boldsymbol{\beta}\|_2} - 2c_\mu \mu \sqrt{\log n} - \frac{4\sqrt{\sigma^2 \log n}}{\|\boldsymbol{\beta}\|_2}} \right\rceil$.

The proof of this theorem is omitted here since it follows in a straightforward manner from Lemma 3.4, Theorem 2.1, and a union bound argument. Nonetheless, it is worth mentioning here that the $k \leq \frac{m}{\log n}$ bound in Lemma 3.4 is omitted in Theorem 3.3 since $k < \frac{\mu^{-2}}{c_\mu^2 \log n} \Rightarrow k \leq \frac{m}{\log n}$ because of the Welch bound. The final result of this section is a corollary of Theorem 3.3 that removes the dependence of d on knowledge of the problem parameters.

Corollary 3. *Let* $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\eta}$ *with* $\boldsymbol{\beta}$ *a* k*-sparse vector and the entries of* $\boldsymbol{\eta}$ *independently distributed as* $\mathcal{N}(0, \sigma^2)$. *Further, assume* X *satisfies the coherence property with*

$c_\mu > 10\sqrt{2}$ *and* $\mathcal{S}$ *is drawn uniformly at random from* k*-subsets of* $[[n]]$*. Finally, suppose* $n \geq \max\{2m, \exp(5)\}$*,* $k < \frac{\mu^{-2}}{c_\mu^2 \log n}$*, and* $\frac{\beta_{\min}}{\|\boldsymbol{\beta}\|_2} > 2c_\mu c_1 \mu \sqrt{\log n} + 4c_2 \frac{\sqrt{\sigma^2 \log n}}{\|\boldsymbol{\beta}\|_2}$ *for some* $c_1, c_2 > 2$*. Then, Algorithm 1 satisfies* $\mathcal{S} \subset \widehat{\mathcal{S}}_d$ *with probability exceeding* $1 - 6n^{-1}$ *as long as* $d \geq \left\lceil \frac{m}{\log n} \right\rceil$*.*

Proof. Since $\frac{\beta_{\min}}{\|\boldsymbol{\beta}\|_2} > 2c_1 c_\mu \mu \sqrt{\log n} + \frac{4c_2\sqrt{\sigma^2 \log n}}{\|\boldsymbol{\beta}\|_2}$, we have that

$$d \geq \left\lceil \frac{\sqrt{k}}{2(c_1 - 1)c_\mu \mu \sqrt{\log n} + \frac{4(c_2 - 1)\sqrt{\sigma^2 \log n}}{\|\boldsymbol{\beta}\|_2}} \right\rceil \tag{3.3}$$

is a sufficient condition for $d \geq \left\lceil \frac{\sqrt{k}}{\frac{\beta_{\min}}{\|\boldsymbol{\beta}\|_2} - 2c_\mu \mu \sqrt{\log n} - \frac{4\sqrt{\sigma^2 \log n}}{\|\boldsymbol{\beta}\|_2}} \right\rceil$. The claim now follows because $d \geq \left\lceil \frac{m}{\log n} \right\rceil$ is a sufficient condition for (3.3), owing to the facts that $n \geq 2m$ and the Welch bound imply $\mu^{-1} \leq \sqrt{2m}$ and $k < \frac{\mu^{-2}}{c_\mu^2 \log n} \Rightarrow k \leq \frac{m}{\log n}$. □

3.3.3 Discussion

Both Theorem 3.2 and Theorem 3.3 shed light on the feasibility of marginal correlation-based screening of linear models *without* imposing a statistical prior on the design matrix. While Theorem 3.2 in this regard provides the least restrictive results, it does suffer from the square-root bottleneck: $k = \mathcal{O}(\mu^{-1}) \Rightarrow k = \mathcal{O}(\sqrt{m})$. Theorem 3.3, on the other hand, overcomes this bottleneck at the expense of uniform prior on the true model as long as $\log n = \mathcal{O}(\mu^{-1})$; in this case, the condition on the sparsity parameter becomes $k = \mathcal{O}\left(\mu^{-2}/\log n\right)$. Therefore, Theorem 3.3 allows for sparsity scaling as high as $k = \mathcal{O}(m/\log n)$ for design matrices with $\mu = \mathcal{O}(m^{-1/2})$; see [56] for existence of such matrices. In addition, Theorem 3.2 and Theorem 3.3 also differ from each other in terms of their respective constraints on $\frac{\beta_{\min}}{\|\boldsymbol{\beta}\|_2}$ for feasibility of marginal correlation-based screening; the constraint in Theorem 3.3 is less restrictive than in Theorem 3.2 for $\log n = \mathcal{O}(\mu^{-1})$.

A natural question to ask at this point is whether Theorem 3.3 specializes to Theorem 3.1. The answer to this question is in the affirmative, except for some small penalties that one has to pay because of the fact that Theorem 3.3 does not exploit any sub-Gaussianity of the entries of X in its analysis. In order to illustrate this further, we

consider the case of Gaussian design matrices and reproduce bounds on their worst-case and average coherences from [56].

Lemma 3.5 ([56, Theorem 8]). *Let $V = [V_{i,j}]$ be an $m \times n$ matrix with the entries $\{V_{i,j}\}_{i,j=1}^{m,n}$ independently distributed as $\mathcal{N}(0,1)$ and $60 \log n \leq m \leq \frac{n-1}{4 \log n}$. Suppose the design matrix X is obtained by normalizing the columns of V, i.e., $X = V\, diag(1/\|V_1\|_2, \ldots, 1/\|V_n\|_2)$. Then, with probability exceeding $1 - 11n^{-1}$, we have*

$$\mu \leq \frac{\sqrt{15 \log n}}{\sqrt{m} - \sqrt{12 \log n}}, \quad \text{and}$$
$$\nu \leq \frac{\sqrt{15 \log n}}{m - \sqrt{12 m \log n}}.$$

It can be seen from this lemma that Gaussian design matrices satisfy the coherence property for $\log n = \mathcal{O}(m^{1/2})$. We can therefore specialize Corollary 3 for Gaussian matrices and conclude that screening of Gaussian linear models using Algorithm 1 can be carried out with $d \geq \left\lceil \frac{m}{\log n} \right\rceil$ as long as: (i) $\log n = \mathcal{O}(m^{1/2})$, (ii) $k = \mathcal{O}(m/(\log n)^2)$, and (iii) $\frac{\beta_{\min}}{\|\boldsymbol{\beta}\|_2} = \Omega\left(\frac{\log n}{\sqrt{m}} + \frac{\sqrt{\sigma^2 \log n}}{\|\boldsymbol{\beta}\|_2}\right)$. Comparing this with Corollary 1 in general and the discussion in Sec. 3.2.2 in particular, we see that the general theory of Sec. 3.3.2 *almost* matches with the specialized theory of Sec. 3.2. Specifically, compared to the constraints of $\log n = \mathcal{O}(m^{1/2})$, $k = \mathcal{O}(m/(\log n)^2)$, and $\frac{\beta_{\min}}{\|\boldsymbol{\beta}\|_2} = \Omega\left(\frac{\log n}{\sqrt{m}} + \frac{\sqrt{\sigma^2 \log n}}{\|\boldsymbol{\beta}\|_2}\right)$ arising in Sec. 3.3.2 for Gaussian design matrices, Sec. 3.2 results in slightly less restrictive constraints of $\log n = \mathcal{O}(m)$, $k = \mathcal{O}(m/\log n)$, and $\frac{\beta_{\min}}{\|\boldsymbol{\beta}\|_2} = \Omega\left(\sqrt{\frac{\log n}{m}} + \frac{\sqrt{\sigma^2 \log n}}{\|\boldsymbol{\beta}\|_2}\right)$. These small gaps are the price one has to pay for the generality of Theorem 3.3.

3.4 Experimental Results

In this section, we present results from a synthetic and a real-data experiment. In Section 3.4.1, we analyze the performance of various regularization-based screening procedures in comparison to the ExSIS procedure. Next, in Section 3.4.2, we analyze the computational savings achieved from the use of ExSIS for screening of the feature space as part of sentiment analysis of IMDb movie reviews [58].

3.4.1 Comparison with Screening Procedures for LASSO-type Methods

In this section, we use Gaussian data to compare the performance of ExSIS to that of screening procedures for LASSO-type methods. The design matrix $X \in \mathbb{R}^{m \times n}$ contains entries from a standard Gaussian distribution such that the pairwise correlation between the variables is ρ. In this experiment, we fix m at 200 while we consider two models with $n = 2000$ and $n = 5000$. For each value of n, we further consider two models with $\rho = 0.0$ and $\rho = 0.3$. Thus, in our experiments, we consider four setups with $(n, \rho) = (2000, 0.0)$, $(2000, 0.3)$, $(5000, 0.0)$ and $(5000, 0.3)$ to analyze the impact of dimensionality and pairwise correlation on performance of the screening procedures for LASSO-type methods in relation to ExSIS. For each of these four different setups, the model size is set at $|\mathcal{S}| = 5$, and the locations of the non-zero coefficients in the parameter vector $\boldsymbol{\beta}$ are chosen such that $\mathcal{S}$ is a uniformly at random subset of $[[n]]$. The values of the non-zero coefficients in the parameter vector $\boldsymbol{\beta}$ are generated from $|z| + 2$ where z is distributed as a standard Gaussian random variable. Furthermore, the noise samples are generated from a standard Gaussian distribution, and the response vector y is generated using (2.1). Finally, the response vector y and the columns of the design matrix X are normalized to have unit norm.

To analyze the performance of screening procedures for the LASSO method [13], the columns of X are screened using SAFE method [43] and strong rules [44] for LASSO. Recall that the LASSO problem can be expressed as

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta} \in \mathbb{R}^n}{\arg\min} \frac{1}{2} \|y - X\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1.$$

For each of these screening methods, we perform screening of X over a set of 200 values of the regularization parameter λ that are chosen uniformly from a linear scale. We compare the screening performance of SAFE method and strong rules for LASSO with the ExSIS method where $d = 2m$. Note that our selection of the value of d has some slack over the suggested value of d from Corollary 1 because the conditions on $\log n$ and $\frac{\beta_{\min}}{\|\boldsymbol{\beta}\|_2}$ in Corollary 1 don't hold true in this experiment.[1] To compare the performance

[1] In order for stated conditions to hold, we need significantly larger m (and n); however, running

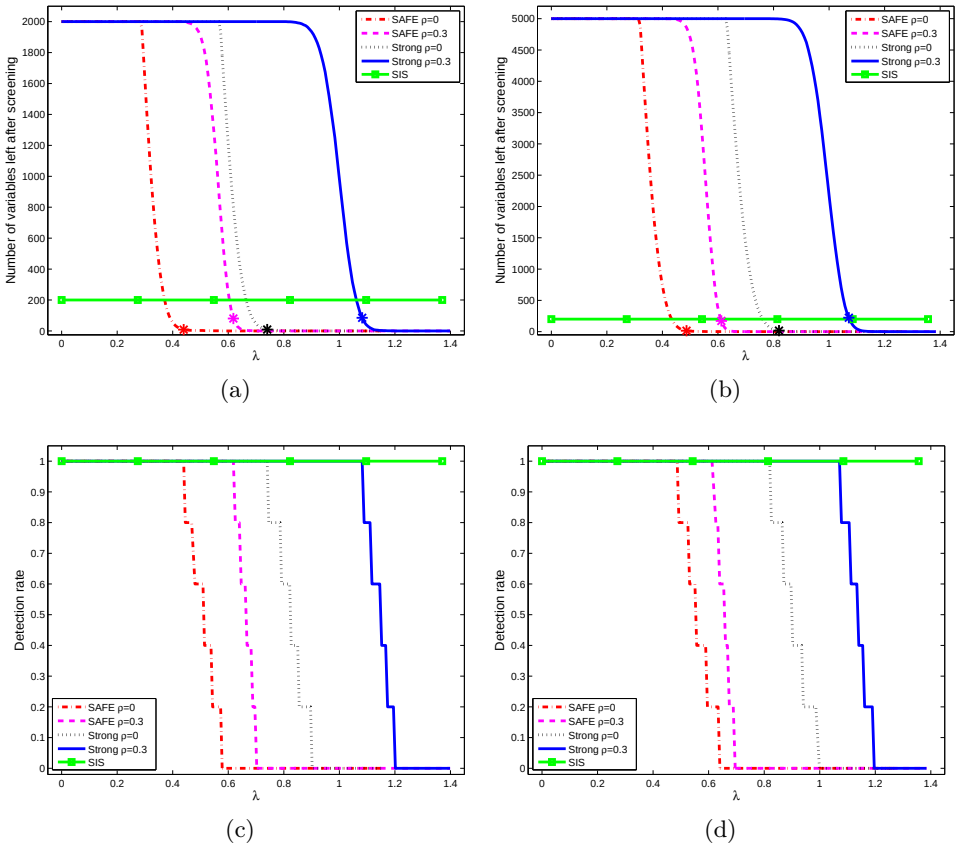


Figure 3.1: Gaussian data matrices are screened using LASSO-based screening procedures for various model parameters. In (a) and (c), $n = 2000$; whereas, in (b) and (d), $n = 5000$. For each value of n, the screening experiment is repeated for $\rho = 0.0$ and $\rho = 0.3$. In each experiment, the model size after screening and the corresponding detection rate is evaluated for different values of the regularization parameter λ. The shown results are median over 100 random draws of the data matrix X/parameter vector $\boldsymbol{\beta}$/noise vector $\boldsymbol{\eta}$ in (2.1).

of these various screening methods, we use two metrics: (i) the model size (number of variables) after screening, which is defined as $|\widehat{\mathcal{S}}_d|$, and (ii) the detection rate, which is defined as $\frac{|\mathcal{S} \cap \widehat{\mathcal{S}}_d|}{|\mathcal{S}|}$. Using these metrics of performance, Fig. 3.1 shows the results of our simulations as median over 100 draws of the random design matrix X/parameter vector $\boldsymbol{\beta}$/noise vector $\boldsymbol{\eta}$ in (2.1) for each of the four setups that we consider in this section.

Next, the design matrix X is also generated and screened using SAFE method [43] and strong rules [44] for elastic net [14] as explained before. Recall that the elastic net

LASSO on such large problems has high computational needs.

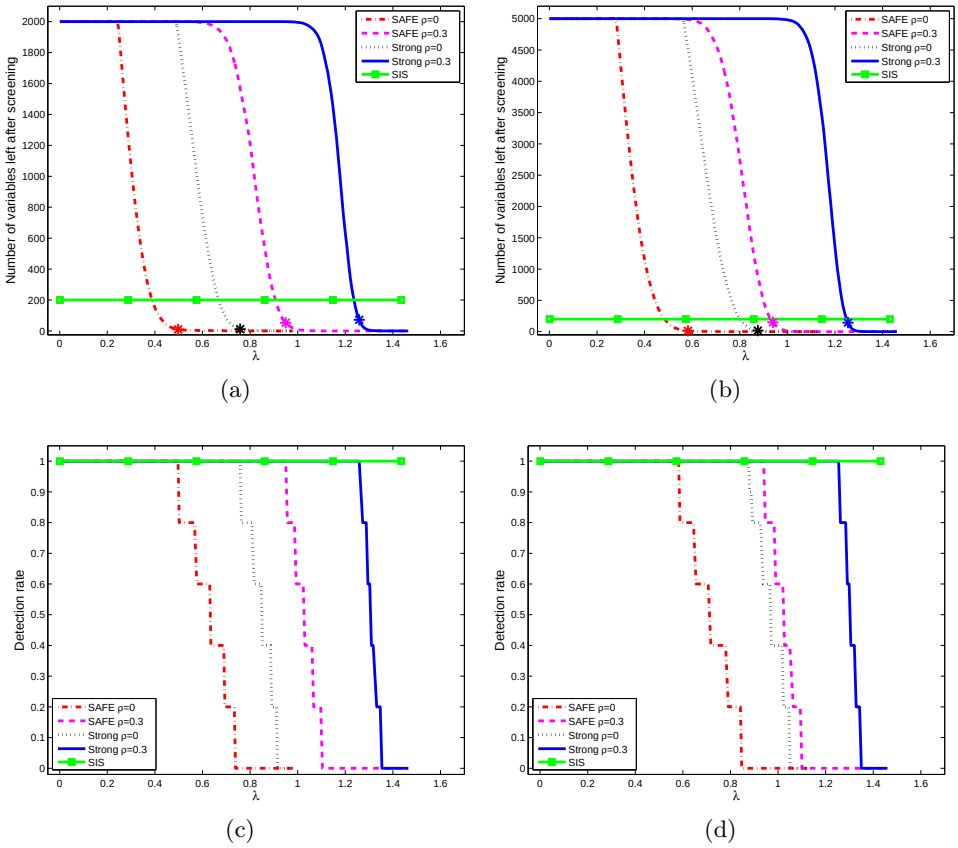


Figure 3.2: Gaussian data matrices are screened using elastic net-based screening procedures for various model parameters. In (a) and (c), $n = 2000$; whereas, in (b) and (d), $n = 5000$. For each value of n, the screening experiment is repeated for $\rho = 0.0$ and $\rho = 0.3$. In each experiment, the model size after screening and the corresponding detection rate is evaluated for different values of the regularization parameter λ. The shown results are median over 100 random draws of the data matrix X/parameter vector $\boldsymbol{\beta}$/noise vector $\boldsymbol{\eta}$ in (2.1).

problem can be expressed as

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^n} \frac{1}{2}\|y - X\boldsymbol{\beta}\|_2^2 + \lambda_1\|\boldsymbol{\beta}\|_1 + \frac{1}{2}\lambda_2\|\boldsymbol{\beta}\|_2.$$

In our simulations, we use the parametrization $(\lambda_1, \lambda_2) = (\alpha\lambda, (1-\alpha)\lambda)$ and we let $\alpha = 0.5$. Fig. 3.2 shows the screening performance of SAFE method for elastic net, strong rules for elastic net and ExSIS over 100 draws of the random design matrix X for each of the four setups, as described before. In both Figs. 3.1 and 3.2, the largest value of λ for which median detection rate is 1.0 is labeled with an asterisk for each optimization-based screening procedure. In other words, for each screening procedure, only if λ is smaller than the value of λ labeled by an asterisk, the screening procedure

maintains a median detection rate of 1.0. Notice also that if the chosen value of λ is too small, no variable is deleted by the screening procedure. Thus, as can be seen in both the figures, there is only a narrow range of values of λ over which the optimization-based screening procedures are able to get rid of variables while maintaining a detection rate of 1.0. Thus, from a practical point of view, it is not trivial to use SAFE method or strong rules for screening because there is no way of ensuring that the chosen value of λ is within the narrow range of values of λ for which significant screening can be performed while maintaining a detection rate of 1.0. In comparison, the ExSIS method does not depend on the parameter λ, and in our experiments, it could always be used for screening while maintaining a median detection rate of 1.0 (as shown in both Figs. 3.1 and 3.2). Before we end this discussion, note that, even within the narrow range of values of λ for which SAFE method or strong rules can be used for screening while maintaining a detection rate of 1.0, there is an even narrower range of values of λ for which SAFE method or strong rules delete more variables than ExSIS.

3.4.2 Sentiment Analysis of IMDb Movie Reviews and ExSIS

In high-dimensional classification, it has been shown that the presence of irrelevant variables increases the difficulty of classification, and the classification error tends to increase with dimensionality of the data model [59, 60]. Variable selection becomes important in high-dimensional classification as it can be used to discard the subset of irrelevant variables and reduce the dimensionality of the data model. Once the variable selection step is performed, classification can be performed based on the subset of relevant variables. In this section, we consider the problem of classifying IMDb movie reviews with positive or negative sentiments. In particular, we use variable selection to (i) reduce dimensionality of the data model, and (ii) learn a linear data model for classification (as explained later). To build and test our classification model, we make use of the IMDb movie reviews dataset [58], with the response being either a 1 (positive review) or a 0 (negative review), and we extract features using the *term frequency-inverse document frequency method* [61].

To increase the reliability of our results, the original dataset of 25K reviews is first

randomly divided into five bins for five independent trials, with each bin further divided into 3K train and 2K test reviews. In each bin, before we use the 3K reviews for fitting a linear model on the feature space, we perform a preprocessing step to get rid of the features (words) that are highly correlated. Note that, when we refer to learning/fitting a linear model, we mean to estimate the vector $\boldsymbol{\beta}$ in (2.1). For learning the linear data model, we use LASSO as well as elastic net. For tuning the regularization parameter in LASSO as well as elastic net for each bin, we perform a five-fold cross validation experiment and choose the value of the regularization parameter that minimizes the mean square error on the training dataset. To evaluate the predictive power of the linear model, we use the notion of test *true positive* (TP) rate, which is the percentage of the remaining 2K test movie reviews that are correctly classified by the model. For classification of the test reviews, we use the trained linear model to estimate the response for each test review. If the estimated response is less than 0.5 for a test review, the test review is assigned a negative sentiment and vice versa. The above procedure is repeated for each of the five bins of data and the average prediction accuracy is reported in Table 3.1. The average model size before variable selection in the five runs of the experiment is $21,345$.

We also repeat the aforementioned experiment procedure but with a slight variation. For each of the five bins of data, we use Algorithm 1 to decrease dimensionality of the data model before performing the variable selection step. The objective of this variation in the experiment is to analyze the decrease in computational time and any change in the prediction accuracy when the variable selection step is preceded with a screening step using Algorithm 1. To choose the value of d in Algorithm 1, we verify that the training data matrix for each fold of data satisfies the coherence property, and then we choose $d = 2m$ where $m = 3000$. Note that the chosen value of d has some slack over the suggested value of d in Corollary 3 because the condition on $\frac{\beta_{\min}}{\|\boldsymbol{\beta}\|_2}$ in Corollary 3 does not hold true in this experiment. After the screening step, we use LASSO and elastic net to learn and test a classification model as explained before. The results are reported in Table 3.1.

Table 3.1: True positive (TP) rates and computational times for experiments on the IMDb dataset, averaged over the five folds of the IMDb dataset. The standard deviation is also reported in parenthesis.

Training method	Train TP rate	Test TP rate	Training time
LASSO	91.35 (0.94)	83.01 (0.77)	388.35 (26.66)
ExSIS-LASSO	98.39 (0.71)	82.23 (0.83)	177.43 (16.85)
Elastic net	96.69 (0.33)	84.35 (0.89)	272.46 (15.76)
ExSIS-Elastic net	99.71 (0.11)	82.06 (0.94)	111.20 (4.65)

Thus, we use LASSO and elastic net, both with and without screening, to train and test a linear model for classification of movie reviews. For each of these four cases, Table 3.1 summarizes the train and test TP rates, which are the percentages of correctly classified reviews in train and test reviews, respectively. The computational time needed for learning the linear model is also reported as an average over the five folds of data. It can be seen from the table that Algorithm 1 reduces the training time by a factor of more than two, while there is only a small decrease in predictive power of the trained model.

3.5 Conclusion

In this chapter, we furthered our understanding of marginal correlation-based screening for ultrahigh-dimensional linear models. In our analysis, we provided verifiable conditions for subGaussian and arbitrary (random or deterministic) linear models under which the dimension of the model can be reduced to almost the sample size. In our experiments with real-world data, we demonstrated the computational savings that can be achieved through ExSIS in high dimensional variable selection.

3.6 Appendix

3.6.1 Proof of Lemma 3.1

Since $X_i := \frac{V_i}{\|V_i\|_2}$, we can write

$$\max_{i\in\mathcal{S}} \big| \sum_{\substack{j\in\mathcal{S}\\ j\neq i}} X_i^\top X_j \boldsymbol{\beta}_j \big| = \max_{i\in\mathcal{S}} \big| \sum_{\substack{j\in\mathcal{S}\\ j\neq i}} \frac{V_i^\top}{\|V_i\|_2} \frac{V_j}{\|V_j\|_2} \boldsymbol{\beta}_j \big|.$$

We next fix an $i' \in \mathcal{S}$ and derive a probabilistic bound on $|\sum_{\substack{j\in\mathcal{S}\\ j\neq i'}} X_{i'}^\top X_j \boldsymbol{\beta}_j|$. This involves deriving both upper and lower probabilistic bounds on $\sum_{\substack{j\in\mathcal{S}\\ j\neq i'}} X_{i'}^\top X_j \boldsymbol{\beta}_j$ below in Step 1 and Step 2, respectively.

Step 1 (Upper Bound): To provide an upper probabilistic bound on $\sum_{\substack{j\in\mathcal{S}\\ j\neq i'}} X_{i'}^\top X_j \boldsymbol{\beta}_j$, we first establish that $\|V_j\|_2 \geq \sqrt{\frac{n\sigma_j^2}{2}}$ for each $j \in \mathcal{S}$ with high probability in Step 1a and then we derive an upper probabilistic bound on $\sum_{\substack{j\in\mathcal{S}\\ j\neq i'}} \frac{V_{i'}^\top}{\|V_{i'}\|_2} \frac{\sqrt{2}V_j}{\sqrt{n\sigma_j^2}} \boldsymbol{\beta}_j$ in Step 1b. We then combine these two steps in Step 1c for the final upper probabilistic bound on $\sum_{\substack{j\in\mathcal{S}\\ j\neq i'}} X_{i'}^\top X_j \boldsymbol{\beta}_j$.

Step 1a: Note that

$$\Pr\left[\|V_j\|_2 < \sqrt{\frac{n\sigma_j^2}{2}}\right] \leq \exp\Big(-\frac{n}{8}\Big(\frac{\sigma_j}{4b_j}\Big)^4\Big) \tag{3.4}$$

for any $j \in \mathcal{S}$ [62, eq. (2.20)]. Next, let $\mathcal{G}_{u,a}$ be the event that $\|V_j\|_2 \geq \sqrt{\frac{n\sigma_j^2}{2}}$ for all $j \in \mathcal{S} \setminus \{i'\}$. Then

$$\begin{aligned}
\Pr[\mathcal{G}_{u,a}^c] &= \Pr\Big[\bigcup_{\substack{j\in\mathcal{S}\\ j\neq i'}} \Big\{\|V_j\|_2 < \sqrt{\frac{n\sigma_j^2}{2}}\Big\}\Big] \\
&\leq \sum_{\substack{j\in\mathcal{S}\\ j\neq i'}} \Pr\left[\|V_j\|_2 < \sqrt{\frac{n\sigma_j^2}{2}}\right] \\
&\leq \sum_{\substack{j\in\mathcal{S}\\ j\neq i'}} \exp\Big(-\frac{n}{8}\Big(\frac{\sigma_j}{4b_j}\Big)^4\Big) \\
&\leq \sum_{\substack{j\in\mathcal{S}\\ j\neq i'}} \exp\Big(-\frac{n}{8}\Big(\frac{\sigma_*}{4b_*}\Big)^4\Big) \\
&= (k-1)\exp\Big(-\frac{n}{8}\Big(\frac{\sigma_*}{4b_*}\Big)^4\Big).
\end{aligned} \tag{3.5}$$

Step 1b: Define

$$Y_{i'} := \sum_{\substack{j\in\mathcal{S}\\ j\neq i'}} \sqrt{\frac{2}{n\sigma_j^2}} \frac{V_{i'}^\top}{\|V_{i'}\|_2} V_j \boldsymbol{\beta}_j,$$

and let $\mathcal{G}_{u,b}$ be the event that $Y_{i'} \leq \sqrt{\frac{8\log p}{n}}\big(\frac{b_*}{\sigma_*}\big)\|\boldsymbol{\beta}\|_2$. Then, the claim is that $\Pr(\mathcal{G}_{u,b}) \geq 1-\frac{1}{p^2}$. In order to prove this claim, let us define another event $\mathcal{G}_v^{'} := \{V_i' = v_{i'}\}$. Then, defining $u_{i'} := \frac{v_{i'}}{\|v_{i'}\|_2}$, we have

$$
\begin{aligned}
M_{Y_{i'}}(\lambda|\,\mathcal{G}_v^{'}) &:= \mathbb{E}[\exp(\lambda Y_{i'})|\,\mathcal{G}_v^{'}] \\
&= \mathbb{E}\Big[\exp\Big(\lambda\sum_{\substack{j\in\mathcal{S}\\ j\neq i'}}\sqrt{\frac{2}{n\sigma_j^2}}\frac{V_{i'}^\top}{\|V_{i'}\|_2}V_j\boldsymbol{\beta}_j\Big)|\,\mathcal{G}_v^{'}\Big] \\
&= \mathbb{E}\Big[\exp\Big(\lambda\sum_{\substack{j\in\mathcal{S}\\ j\neq i'}}\sqrt{\frac{2}{n\sigma_j^2}}u_{i'}^\top V_j\boldsymbol{\beta}_j\Big)\Big] \\
&= \mathbb{E}\Big[\prod_{\substack{j\in\mathcal{S}\\ j\neq i'}}\exp\Big(\lambda\sqrt{\frac{2}{n\sigma_j^2}}u_{i'}^\top V_j\boldsymbol{\beta}_j\Big)\Big] \\
&= \mathbb{E}\Big[\prod_{\substack{j\in\mathcal{S}\\ j\neq i'}}\exp\Big(\lambda\sqrt{\frac{2}{n\sigma_j^2}}\sum_{l=1}^{n}u_{i',l}V_{j,l}\boldsymbol{\beta}_j\Big)\Big] \\
&= \mathbb{E}\Big[\prod_{\substack{j\in\mathcal{S}\\ j\neq i'}}\prod_{l=1}^{n}\exp\Big(\lambda\sqrt{\frac{2}{n\sigma_j^2}}u_{i',l}V_{j,l}\boldsymbol{\beta}_j\Big)\Big] \\
&= \prod_{\substack{j\in\mathcal{S}\\ j\neq i'}}\prod_{l=1}^{n}\mathbb{E}\Big[\exp\Big(\lambda\sqrt{\frac{2}{n\sigma_j^2}}u_{i',l}V_{j,l}\boldsymbol{\beta}_j\Big)\Big] \\
&\leq \prod_{\substack{j\in\mathcal{S}\\ j\neq i'}}\prod_{l=1}^{n}\exp\Big(\frac{\lambda^2}{n\sigma_j^2}u_{i',l}^2b_j^2\boldsymbol{\beta}_j^2\Big) \\
&= \prod_{\substack{j\in\mathcal{S}\\ j\neq i'}}\exp\Big(\frac{\lambda^2}{n\sigma_j^2}b_j^2\boldsymbol{\beta}_j^2\sum_{l=1}^{n}u_{i',l}^2\Big) \\
&= \prod_{\substack{j\in\mathcal{S}\\ j\neq i'}}\exp\Big(\frac{\lambda^2}{n\sigma_j^2}b_j^2\boldsymbol{\beta}_j^2\Big) \\
&= \exp\Big(\frac{\lambda^2}{n}\sum_{\substack{j\in\mathcal{S}\\ j\neq i'}}\frac{b_j^2\boldsymbol{\beta}_j^2}{\sigma_j^2}\Big) \\
&\leq \exp\Big(\frac{\lambda^2}{n}\Big(\frac{b_*}{\sigma_*}\Big)^2\sum_{\substack{j\in\mathcal{S}\\ j\neq i'}}\boldsymbol{\beta}_j^2\Big)
\end{aligned}
$$

$$\leq \exp\Big(\frac{\lambda^2}{n}\Big(\frac{b_*}{\sigma_*}\Big)^2\|\boldsymbol{\beta}\|_2^2\Big). \tag{3.6}$$

By the Chernoff bound on $Y_{i'}$, we next obtain

$$\begin{aligned}\Pr(Y_{i'} > a|\,\mathcal{G}_v^{'}) &\leq \min_{\lambda>0}\exp(-\lambda a)M_{Y_{i'}}(\lambda|\,\mathcal{G}_v^{'})\\ &\leq \min_{\lambda>0}\exp(-\lambda a)\exp\Big(\frac{\lambda^2}{n}\Big(\frac{b_*}{\sigma_*}\Big)^2\|\boldsymbol{\beta}\|_2^2\Big)\\ &= \exp\Big(-\frac{1}{4}\Big(\frac{\sigma_*}{b_*}\Big)^2\frac{a^2\,n}{\|\boldsymbol{\beta}\|_2^2}\Big).\end{aligned} \tag{3.7}$$

Substituting $a = \sqrt{\frac{8\log p}{n}}\Big(\frac{b_*}{\sigma_*}\Big)\|\boldsymbol{\beta}\|_2$ in (3.7), we obtain

$$\Pr\left(Y_{i'} > \sqrt{\frac{8\log p}{n}}\Big(\frac{b_*}{\sigma_*}\Big)\|\boldsymbol{\beta}\|_2\Big|\,\mathcal{G}_v^{'}\right) \leq \frac{1}{p^2}. \tag{3.8}$$

Finally, by the law of total probability, we obtain

$$\begin{aligned}\Pr\left(Y_{i'} > \sqrt{\frac{8\log p}{n}}\Big(\frac{b_*}{\sigma_*}\Big)\|\boldsymbol{\beta}\|_2\right) &= \mathbb{E}_{V_i'}\left[\Pr\left(Y_{i'} > \sqrt{\frac{8\log p}{n}}\Big(\frac{b_*}{\sigma_*}\Big)\|\boldsymbol{\beta}\|_2\Big|\,\mathcal{G}_v^{'}\right)\right]\\ &\leq \mathbb{E}_{V_i'}\left[\frac{1}{p^2}\right] = \frac{1}{p^2}.\end{aligned} \tag{3.9}$$

Thus, the event $\mathcal{G}_{u,b}$ holds with probability exceeding $1-\frac{1}{p^2}$.

Step 1c: Conditioning on $\mathcal{G}_{u,a}\cap\mathcal{G}_{u,b}$, we have from (3.5) and (3.9) that

$$\sum_{\substack{j\in\mathcal{S}\\ j\neq i'}} X_{i'}^\top X_j\boldsymbol{\beta}_j \leq \sqrt{\frac{8\log p}{n}}\Big(\frac{b_*}{\sigma_*}\Big)\|\boldsymbol{\beta}\|_2. \tag{3.10}$$

Further, note that

$$\begin{aligned}\Pr(\mathcal{G}_{u,a}\cap\mathcal{G}_{u,b}) &\geq 1-\Pr(\mathcal{G}_{u,a}^c)-\Pr(\mathcal{G}_{u,b}^c)\\ &\geq 1-(k-1)\exp\Big(-\frac{n}{8}\Big(\frac{\sigma_*}{4b_*}\Big)^4\Big)-\frac{1}{p^2}\\ &\overset{\text{(a)}}{\geq} 1-\frac{(k-1)}{p^2}-\frac{1}{p^2} = 1-\frac{k}{p^2},\end{aligned}$$

where (a) follows since $\log p \leq \frac{n}{16}\big(\frac{\sigma_*}{4b_*}\big)^4$. Thus, (3.10) holds with probability exceeding $1-\frac{k}{p^2}$.

Step 2 (Lower Bound): Our claim in this step is that $\sum\limits_{\substack{j\in\mathcal{S}\\ j\neq i'}} X_{i'}^\top X_j\boldsymbol{\beta}_j \geq -\sqrt{\frac{8\log p}{n}}\Big(\frac{b_*}{\sigma_*}\Big)\|\boldsymbol{\beta}\|_2$

with probability exceeding $1 - \frac{k}{p^2}$. To establish this claim, notice that

$$\sum_{\substack{j\in\mathcal{S}\\ j\neq i'}} X_{i'}^\top X_j\boldsymbol{\beta}_j \geq -\sqrt{\frac{8\log p}{n}}\Big(\frac{b_*}{\sigma_*}\Big)\|\boldsymbol{\beta}\|_2 \quad \Longleftrightarrow \quad -\sum_{\substack{j\in\mathcal{S}\\ j\neq i'}} X_{i'}^\top X_j\boldsymbol{\beta}_j \leq \sqrt{\frac{8\log p}{n}}\Big(\frac{b_*}{\sigma_*}\Big)\|\boldsymbol{\beta}\|_2. \tag{3.11}$$

Further, we have $-\sum_{\substack{j\in\mathcal{S}\\ j\neq i'}} X_{i'}^\top X_j\boldsymbol{\beta}_j \equiv -\sum_{\substack{j\in\mathcal{S}\\ j\neq i'}} \frac{V_{i'}^\top}{\|V_{i'}\|_2}\frac{V_j}{\|V_j\|_2}\boldsymbol{\beta}_j = \sum_{\substack{j\in\mathcal{S}\\ j\neq i'}} \frac{V_{i'}^\top}{\|V_{i'}\|_2}\frac{\widetilde{V}_j}{\|V_j\|_2}\boldsymbol{\beta}_j$, where $\widetilde{V}_j := -V_j$ is still distributed as V_j because of the symmetry of sub-Gaussian distributions. The claim now follows from a repetition of the analysis carried out in Step 1.

Final Step: Step 1 and Step 2, along with the union bound, imply that

$$|\sum_{\substack{j\in\mathcal{S}\\ j\neq i}} X_{i'}^\top X_j\boldsymbol{\beta}_j| \leq \sqrt{\frac{8\log p}{n}}\Big(\frac{b_*}{\sigma_*}\Big)\|\boldsymbol{\beta}\|_2$$

with probability exceeding $1 - \frac{2k}{p^2}$. Next, notice that

$$\begin{aligned}
&\Pr\Big[\max_{i\in\mathcal{S}}|\sum_{\substack{j\in\mathcal{S}\\ j\neq i}} X_i^\top X_j\boldsymbol{\beta}_j| \leq \sqrt{\frac{8\log p}{n}}\Big(\frac{b_*}{\sigma_*}\Big)\|\boldsymbol{\beta}\|_2\Big]\\
&= 1 - \Pr\Big[\bigcup_{i\in\mathcal{S}}\Big\{|\sum_{\substack{j\in\mathcal{S}\\ j\neq i}} X_i^\top X_j\boldsymbol{\beta}_j| > \sqrt{\frac{8\log p}{n}}\Big(\frac{b_*}{\sigma_*}\Big)\|\boldsymbol{\beta}\|_2\Big\}\Big]\\
&\geq 1 - \sum_{i\in\mathcal{S}}\Pr\Big[|\sum_{\substack{j\in\mathcal{S}\\ j\neq i}} X_i^\top X_j\boldsymbol{\beta}_j| > \sqrt{\frac{8\log p}{n}}\Big(\frac{b_*}{\sigma_*}\Big)\|\boldsymbol{\beta}\|_2\Big]\\
&\geq 1 - \sum_{i\in\mathcal{S}}\frac{2k}{p^2} = 1 - \frac{2k^2}{p^2}.
\end{aligned} \tag{3.12}$$

This complete the proof of the lemma. □

3.6.2 Proof of Lemma 3.2

Once again, notice that

$$\max_{i\in\mathcal{S}^c}|\sum_{j\in\mathcal{S}} X_i^\top X_j\boldsymbol{\beta}_j| = \max_{i\in\mathcal{S}^c}\Big|\sum_{j\in\mathcal{S}}\frac{V_i^\top}{\|V_i\|_2}\frac{V_j}{\|V_j\|_2}\boldsymbol{\beta}_j\Big|.$$

We next fix an $i' \in \mathcal{S}^c$. Similar to the proof of Lemma 3.1, the plan is to first derive a probabilistic bound on $|\sum_{j\in\mathcal{S}} X_{i'}^\top X_j\boldsymbol{\beta}_j|$ and then use the union bound to provide a

probabilistic bound on $\max_{i\in\mathcal{S}^c}|\sum_{j\in\mathcal{S}} X_i^\top X_j\boldsymbol{\beta}_j|$. Using steps similar to the ones in the proof of Lemma 3.1, it is straightforward to establish that

$$|\sum_{j\in\mathcal{S}} X_{i'}^\top X_j\boldsymbol{\beta}_j| \leq \sqrt{\frac{8\log p}{n}}\Big(\frac{b_*}{\sigma_*}\Big)\|\boldsymbol{\beta}\|_2$$

with probability exceeding $1-\frac{2(k+1)}{p^2}$. The union bound finally gives us

$$\begin{aligned}
&\Pr\Big[\max_{i\in\mathcal{S}^c}|\sum_{j\in\mathcal{S}} X_i^\top X_j\boldsymbol{\beta}_j| \leq \sqrt{\frac{8\log p}{n}}\Big(\frac{b_*}{\sigma_*}\Big)\|\boldsymbol{\beta}\|_2\Big] \\
&= 1-\Pr\Big[\bigcup_{i\in\mathcal{S}^c}\Big\{|\sum_{j\in\mathcal{S}} X_i^\top X_j\boldsymbol{\beta}_j| > \sqrt{\frac{8\log p}{n}}\Big(\frac{b_*}{\sigma_*}\Big)\|\boldsymbol{\beta}\|_2\Big\}\Big] \\
&\geq 1-\sum_{i\in\mathcal{S}^c}\Pr\Big[|\sum_{j\in\mathcal{S}} X_i^\top X_j\boldsymbol{\beta}_j| > \sqrt{\frac{8\log p}{n}}\Big(\frac{b_*}{\sigma_*}\Big)\|\boldsymbol{\beta}\|_2\Big] \\
&\geq 1-\sum_{i\in\mathcal{S}^c}\frac{2(k+1)}{p^2} = 1-\frac{2(k+1)(p-k)}{p^2}. \qquad (3.13)
\end{aligned}$$

This completes the proof of the lemma. □

3.6.3 Proof of Lemma 3.3

Notice that $\max_{i\in\mathcal{S}}|\sum_{\substack{j\in\mathcal{S}\\ j\neq i}} X_i^\top X_j\boldsymbol{\beta}_j| \leq \max_{i\in\mathcal{S}}\sum_{\substack{j\in\mathcal{S}\\ j\neq i}}|X_i^\top X_j\boldsymbol{\beta}_j| \leq \max_{i\in\mathcal{S}}\sum_{\substack{j\in\mathcal{S}\\ j\neq i}}|X_i^\top X_j||\boldsymbol{\beta}_j|$. Further, we have

$$\max_{i\in\mathcal{S}}\sum_{\substack{j\in\mathcal{S}\\ j\neq i}}|X_i^\top X_j||\boldsymbol{\beta}_j| \leq \max_{i\in\mathcal{S}}\sum_{\substack{j\in\mathcal{S}\\ j\neq i}}\mu|\boldsymbol{\beta}_j| \leq \mu\|\boldsymbol{\beta}\|_1 \leq \mu\sqrt{k}\|\boldsymbol{\beta}\|_2. \qquad (3.14)$$

An identical argument also establishes that $\max_{i\in\mathcal{S}^c}|\sum_{j\in\mathcal{S}} X_i^\top X_j\boldsymbol{\beta}_j| \leq \mu\|\boldsymbol{\beta}\|_1 \leq \mu\sqrt{k}\|\boldsymbol{\beta}\|_2$.

□

3.6.4 Proof of Lemma 3.4

The proof of Lemma 3.4 relies on the following two lemmas, which are formally proved in [50].

Lemma 3.6 ([50, Lemma 3]). *Assume $\mathcal{S}$ is drawn uniformly at random from k-subsets of $[[p]]$ and choose a parameter $a \geq 1$. Let $\epsilon \in [0,1)$ and $k \leq \min\{\epsilon^2\nu^{-2}, (1+a)^{-1}p\}$.*

Then, with probability exceeding $1 - 4k\exp(-\frac{(\epsilon-\sqrt{k}\nu)^2}{16(2+a^{-1})^2\mu^2})$*, we have*

$$\max_{i\in\mathcal{S}} |\sum_{\substack{j\in\mathcal{S}\\ j\neq i}} X_i^\top X_j\boldsymbol{\beta}_j| \leq \epsilon \|\boldsymbol{\beta}\|_2.$$

Lemma 3.7 [50, Lemma 4])**.** *Assume* $\mathcal{S}$ *is drawn uniformly at random from* k*-subsets of* $[[p]]$ *and choose a parameter* $a \geq 1$*. Let* $\epsilon \in [0, 1)$ *and* $k \leq \min\{\epsilon^2\nu^{-2}, (1+a)^{-1}p\}$*. Then, with probability exceeding* $1 - 4(p-k)\exp(-\frac{(\epsilon-\sqrt{k}\nu)^2}{8(2+a^{-1})^2\mu^2})$*, we have*

$$\max_{i\in\mathcal{S}^c} |\sum_{j\in\mathcal{S}} X_i^\top X_j\boldsymbol{\beta}_j| \leq \epsilon \|\boldsymbol{\beta}\|_2.$$

Using a simple union bound with Lemma 3.6 and Lemma 3.7, we have

$$\max_{i\in\mathcal{S}} |\sum_{\substack{j\in\mathcal{S}\\ j\neq i}} X_i^\top X_j\boldsymbol{\beta}_j| \leq \epsilon \|\boldsymbol{\beta}\|_2, \text{ and}$$

$$\max_{i\in\mathcal{S}^c} |\sum_{j\in\mathcal{S}} X_i^\top X_j\boldsymbol{\beta}_j| \leq \epsilon \|\boldsymbol{\beta}\|_2$$

with probability exceeding $1 - \delta$ where $\delta = 4p\exp(-\frac{(\epsilon-\sqrt{k}\nu)^2}{16(2+a^{-1})^2\mu^2})$. Fix $\epsilon = c_\mu\mu\sqrt{\log p}$ and $a = 9$. Then, the claim is that $\delta \leq 4p^{-1}$. Next, we will prove our claim. Before we can fix $\epsilon = c_\mu\mu\sqrt{\log p}$ and $a = 9$, we need to ensure that the chosen values of a, ϵ and the allowed values of k satisfy the assumptions in Lemma 3.6 and Lemma 3.7. First, note that $\epsilon < 1$ because of $\mu < \frac{1}{c_\mu\sqrt{\log p}}$. Second, $k \leq \frac{p}{1+a}$ because of $a = 9$, $p \geq \max\{2n, \exp(5)\}$ and $k \leq \frac{n}{\log p}$. Third, and last, $k \leq \frac{\epsilon^2}{(9\nu)^2}$ because of $\nu < \frac{\mu}{\sqrt{n}}$, $k \leq \frac{n}{\log p}$, $p \geq 2n$ and $c_\mu > 10\sqrt{2}$. Finally, we use $\sqrt{k}\nu \leq \frac{\epsilon}{9}$ with $a = 9$, $c_\mu > 10\sqrt{2}$ and $\epsilon = c_\mu\mu\sqrt{\log p}$ to obtain $\exp(-\frac{(\epsilon-\sqrt{k}\nu)^2}{16(2+a^{-1})^2\mu^2}) \leq p^{-2}$ and thus $\delta \leq 4p^{-1}$. □

Chapter 4

Linear Tensor Regression Model

In this chapter, we study a new regression model with a scalar response variable and tensor-valued predictors, where the parameters form a tensor in $\mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d}$, and the parameter tensor simultaneously embeds structures of low rank and sparsity. Specifically, we focus on the task of estimating the parameter tensor from m observations of the response variable and the tensor-valued predictors. We formulate parameter estimation as a non-convex problem, and we propose a projected gradient descent-based method to solve it. We also provide theoretical guarantees for the proposed method, in which we show that the method converges to the correct solution within a certain number of iterations, based on a certain Restricted Isometry Property. In our experiments on synthetic data, we demonstrate the efficacy of the proposed learning method for learning the proposed regression model.

4.1 Introduction

In this chapter, we consider the ordinary linear model in (1.1) for the case when $d \geq 3$. A major application of this model can be found in neuroimaging data analysis, where the voxels (predictors) in a brain image naturally appear in the form of a tensor and the associated disease outcome (response) appears as a scalar variable [8, 9, 63, 64]. Given $\{\mathbf{X}_i\}_{i=1}^m$ and $\{y_i\}_{i=1}^m$ in (1.1), we focus on the task of learning the regression model in (1.1), which is equivalent to estimating $\mathbf{B}$. One simple approach for estimating $\mathbf{B}$ is to vectorize the tensors, and then use any of the traditional sparsity promoting methods for learning the regresison model. Specifically, the parameter tensor $\mathbf{B}$ and the predictor tensors $\{\mathbf{X}_i\}_{i=1}^m$ can be vectorized such that the model in (1.1) can equivalently be expressed as $y_i = \langle \text{vec}(\mathbf{X}_i), \text{vec}(\mathbf{B}) \rangle + \eta_i$, where vec(.) denotes the vectorization

procedure. Given this vector-valued regression model, any of the traditional sparsity promoting techniques in the literature—such as forward selection/matching pursuit, backward elimination [12], least absolute shrinkage and selection operator (LASSO) [13], elastic net [14], bridge regression [30, 31], adaptive LASSO [32], group LASSO [33], and Dantzig selector [34]—can be employed for estimating $\text{vec}(\mathbf{B})$. However, typically in tensor regression models, the number of predictors are massive compared with the sample size, and thus the application of the aforementioned learning techniques can easily become computationally prohibitive. To address this computational bottleneck, various two-stage approaches have been proposed in the literature, where the learning stage is preceded by application of a dimensionality reduction step to reduce the number of predictors in the model. Some prominent examples of such dimensionality reduction methods include principal component analysis [65] and variable screening [36, 66, 67] methods.

Besides the computational bottleneck of vectorization-based learning approaches, another major drawback of vectorization is that the spatial structure among the entries of the tensor $\mathbf{B}$ is not preserved—structure that can possibly be exploited for efficient estimation of $\mathbf{B}$. Among the various tensor decompositions that capture such spatial relationships among tensor entries [16, 20], a popular decomposition is the Tucker decomposition. The notion of rank associated with Tucker decomposition of a tensor $\mathbf{B}$ is known as the Tucker rank, which is an n-dimensional tuple whose i-th entry is the column rank of the mode-i unfolding $\mathbf{B}_{(i)}$ of $\mathbf{B}$:

$$\big(\text{rank}(\mathbf{B}_{(1)}), \text{rank}(\mathbf{B}_{(2)}), \ldots, \text{rank}(\mathbf{B}_{(d)})\big).$$

This notion of Tucker rank has been successfully employed for learning tensor-valued regression models [10, 21, 22] under the imposition of low Tucker rank on $\mathbf{B}$. Specifically, some early approaches were based on minimization of the sum of nuclear norm of matricizations of tensor $\mathbf{B}$ in each mode [10, 21, 23, 68]. To analyze the sample complexity of learning methods in the literature, suppose $(r, r, \ldots, r)$ is the Tucker rank of $\mathbf{B}$, $n := n_1 = n_2 = \cdots = n_d$, and $\mathbf{X}_i$ draws values from Gaussian distribution for $i \in [[m]]$. Under these suppositions, it was shown that approaches based on sum of nuclear norm

minimization require $\Omega(rn^{(d-1)})$ samples [23], which is highly suboptimal. Thus, more recently, tensor variants of the projected gradient method [69–71] have gained popularity for solving non-convex formulations of the learning problem [22, 72, 73]. In such recent work that studies the imposition of low Tucker rank on $\mathbf{B}$ [22], it was shown that $\mathbf{B}$ can be learnt using $\mathcal{O}\big((r^d + nrd)\log d\big)$ observations. Note that this sample complexity bound is order optimal up to a logarithmic factor.

Although consideration of low Tucker rank allows for efficient learning, the sample complexity requirement still has a linear dependence on n, since the degrees of freedom of $\mathbf{B}$ are $\mathcal{O}(r^d + nrd)$. Such sample complexity requirement can become prohibitive in application areas like neuroimaging data analysis. For example, given a typical MRI image of size $256 \times 256 \times 256$ with $r = 3$ and $d = 3$, we have $nrd = 1152$; whereas, the number of subjects are typically less than ≈ 1000 [24]. Thus, an interesting research question is whether we can remove the linear dependence of the sample complexity on n, while still being able to exploit the spatial relationships among the entries of the tensor parameter? Clearly, the sample complexity dependence on n is unavoidable in Tucker decomposition, because the degrees of freedom in the Tucker model scale linearly with n. Another problem with Tucker decomposition is that of model interpretability: unlike the imposition of sparsity on the parameters by the application of traditional variable selection methods [12–14], the parameters are typically non-zero with the imposition of low Tucker rank on $\mathbf{B}$.

We address these shortcomings of the Tucker decomposition by considering a tensor structure that massively reduces the degrees of freedom in $\mathbf{B}$ while increasing interpretability of the regression model in (1.1). Specifically, we study the simultaneous imposition of low Tucker rank and sparsity on $\mathbf{B}$, where sparsity on $\mathbf{B}$ is invoked by imposing sparsity on the factor matrices of Tucker decomposition. In the next section, we formally motivate and define the simultaneous imposition of low rank and sparsity on the parameter tensor, and then we explicitly outline our contributions.

4.2 Model Setup

4.2.1 Background on Tensor Decompositions

The Tucker decomposition [16, 20] decomposes a tensor into a core tensor transformed by different factor matrices along different modes. Using Tucker decomposition, any arbitrary tensor $\mathbf{B} \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d}$ can be written as

$$\mathbf{B} = \widetilde{\mathbf{G}} \times_1 \widetilde{U}_1 \times_2 \widetilde{U}_2 \cdots \times_d \widetilde{U}_d, \tag{4.1}$$

where $\widetilde{\mathbf{G}} \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d}$ is the core tensor, and $\widetilde{U}_i \in \mathbb{R}^{n_i \times n_i}$, $i \in [[d]]$, are the factor matrices. Let $\mathbf{B}_{(i)}$ be the mode-i matricization of $\mathbf{B}$, such that the columns of $\mathbf{B}_{(i)}$ contain the mode-i fibers of $\mathbf{B}$. Then, the mode-i matricization of $\mathbf{B}$ can be expressed as

$$\mathbf{B}_{(i)} = \widetilde{U}_i \, \widetilde{\mathbf{G}}_{(i)} (\widetilde{U}_n \otimes \cdots \otimes \widetilde{U}_{(i+1)} \otimes \widetilde{U}_{(i-1)} \otimes \cdots \otimes \widetilde{U}_1)^\top, \tag{4.2}$$

where $\widetilde{\mathbf{G}}_{(i)} \in \mathbb{R}^{n_i \times \prod_{j \neq i} n_j}$ is the mode-i matricization of the core tensor $\widetilde{\mathbf{G}}$. Let r_i be the column rank of $\mathbf{B}_{(i)}$, and let $U_i \in \mathbb{R}^{n_i \times r_i}$ be a basis of the column span of $\mathbf{B}_{(i)}$. Multiplying both sides of (4.2) by $U_i (U_i^\top U_i)^{-1} U_i^\top$, which is a projection matrix for the column space of $\mathbf{B}_{(i)}$, we obtain

$$\mathbf{B}_{(i)} = U_i (U_i^\top U_i)^{-1} U_i^\top \, \widetilde{U}_i \, \widetilde{\mathbf{G}}_{(i)} (\widetilde{U}_n \otimes \cdots \otimes \widetilde{U}_{(i+1)} \otimes \widetilde{U}_{(i-1)} \otimes \cdots \otimes \widetilde{U}_1)^\top, \tag{4.3}$$

Then, without loss of generality, we can absorb the linear transformation $(U_i^\top U_i)^{-1} U_i^\top \widetilde{U}_i$ into $\widetilde{\mathbf{G}}_{(i)}$. Correspondingly, $\widetilde{\mathbf{G}}_{(i)}$ is transformed from $\mathbb{R}^{n_i \times \prod_{j \neq i} n_j}$ to $\mathbb{R}^{r_i \times \prod_{j \neq i} n_j}$, and (4.3) can be expressed as

$$\mathbf{B}_{(i)} = U_i \, \widetilde{\mathbf{G}}_{(i)} (\widetilde{U}_n \otimes \cdots \otimes \widetilde{U}_{(i+1)} \otimes \widetilde{U}_{(i-1)} \otimes \cdots \otimes \widetilde{U}_1)^\top, \tag{4.4}$$

where $U_i \in \mathbb{R}^{n_i \times r_i}$ and $\widetilde{\mathbf{G}}_{(i)} \in \mathbb{R}^{r_i \times \prod_{j \neq i} n_j}$. Carrying out these transformations from (4.2) to (4.4) for all modes $i \in [[d]]$, the tensor $\mathbf{B}$, without loss of generality, can be expressed as

$$\mathbf{B} = \mathbf{G} \times_1 U_1 \times_2 U_2 \cdots \times_d U_d, \tag{4.5}$$

where $\mathbf{G} \in \mathbb{R}^{r_1 \times r_2 \times \cdots \times r_d}$ is the core tensor, $U_i \in \mathbb{R}^{n_i \times r_i}$, $i \in [[d]]$, are the factor matrices, and r_i is the column-rank of $\mathbf{B}_{(i)}$, $i \in [[d]]$.

4.3 Problem Formulation

For ease of notation, let us define $\mathcal{W} := \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d}$, $\mathcal{Y} := \mathbb{R}^m$, and let us denote the collection of tensors $\{\mathbf{X}_i\}_{i=1}^m$ in (1.1) by a linear map/measurement operator $\mathcal{X} : \mathcal{W} \to \mathcal{Y}$ such that (1.1) can equivalently be expressed as

$$\mathbf{y} = \mathcal{X}(\mathbf{B}) + \boldsymbol{\eta}, \tag{4.6}$$

where $\mathbf{y} = [y_1, y_2, \ldots, y_m]$, and $\boldsymbol{\eta} = [\eta_1, \eta_2, \ldots, \eta_m]$. In this work, we impose that the parameter tensor $\mathbf{B} \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d}$ is structured in the sense that it is $\underline{r}$-rank and $\underline{s}$-sparse, where the notion of an $\underline{r}$-rank and $\underline{s}$-sparse tensor is defined as follows.

Definition 4.1 ($\underline{r}$-rank and $\underline{s}$-sparse tensor)**.** Given a rank tuple $\underline{r} := (r_1, r_2, \ldots, r_d)$ and a sparsity tuple $\underline{s} := (s_1, s_2, \ldots, s_d)$, a tensor $\mathbf{Z} \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d}$ is said to be $\underline{r}$-rank and $\underline{s}$-sparse if $\mathbf{Z}$ can be expressed as

$$\mathbf{Z} = \mathbf{S} \times_1 U_1 \times_2 U_2 \cdots \times_d U_d, \tag{4.7}$$

where $\mathbf{S} \in \mathbb{R}^{r_1 \times r_2 \times \cdots \times r_d}$ and $U_i \in \mathbb{R}^{n_i \times r_i}$, with $\|U_i(:, j)\|_0 \leq s_i$, $\forall i \in [[d]]$, $j \in [[r_i]]$. Notice that, trivially, $r_i \leq n_i$ and $s_i \leq n_i$.

Recall from [16] that (4.7) is expressing $\mathbf{Z}$ in terms of a Tucker decomposition, in which $\mathbf{S}$ is termed the core tensor and the U_i's are referred to as factor matrices, with additional sparsity constraints on the factor matrices. It can also be seen from (4.7) that for the special case when $s_i = n_i$, the mode-i matricization of $\mathbf{Z}$ has rank r_i: $\text{rank}(\mathbf{Z}_{(i)}) = \underline{r}_i$; i.e., the $\underline{r}$-rank of $\mathbf{Z}$ is simple the Tucker rank of $\mathbf{Z}$. Further, note that we are defining sparsity of $\mathbf{Z}$ in terms of sparsity of the columns of the factor matrices $\{U_i(:, j)\}$, $i \in [[d]]$, $j \in r_i$, that are generating the tensor. This notion of sparsity is different from the conventional notion of sparsity, where sparsity is defined as the number of non-zero entries for the data structure under consideration, i.e., tensor $\mathbf{Z}$ in this case. In contrast, the said notion of sparsity not only induces sparsity on $\mathbf{Z}$ but also dramatically reduces the number of free parameters in $\mathbf{Z} \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d}$ from $n := \prod_i n_i$ to the order of $\prod_i r_i + \sum_i r_i s_i \log n_i$, which can be significantly smaller than n for $\underline{r}_i \ll n_i$ and $s_i \ll n_i$ (the $\log n_i$ factor arises from the need to encode the locations of the s_i non-zero entries

in a given column of U_i). This reduction in degrees of freedom allows us to learn the tensor regression model in (4.6) with lower sample complexity, as we show later.

Since we are imposing the unknown tensor $\mathbf{B}$ is $\underline{r}$-rank and $\underline{s}$-sparse in our regression model (4.6), we formally define a set of such tensors as follows:

$$\begin{aligned}\mathcal{C} =&\{\mathbf{S} \times_1 U_1 \times_2 U_2 \times_3 \cdots \times_d U_d : \mathbf{S} \in \mathbb{R}^{r_1 \times r_2 \times \cdots \times r_d}, \text{ and} \\ & U_i \in \mathbb{R}^{n_i \times r_i}, \|U_i(:,j)\|_0 \leq s_i, i \in [[d]], j \in [[r_i]]\}. \end{aligned} \tag{4.8}$$

Using the definition of constraint set $\mathcal{C}$, and given a known linear map $\mathcal{X}$, we can pose the following constrained optimization problem for recovery of $\mathbf{B}$ from noisy linear measurements $\mathbf{y}$:

$$\hat{\mathbf{B}} = \arg\min_{\mathbf{Z} \in \mathcal{C}} \frac{1}{2} \|\mathbf{y} - \mathcal{X}(\mathbf{Z})\|_2^2. \tag{4.9}$$

We can see that the optimization problem posed in (4.9) is non-convex because of non-convexity of the constraint set $\mathcal{C}$. In contrast, most of the prior works in tensor parameter estimation focus on solving convex relaxations of the tensor recovery problem for various notions of low-dimensional tensor structures [10, 21, 23, 68], hence benefiting from rich literature on theory and algorithms for convex optimization. But the issue with convex relaxation-based solutions is that *convex relaxations can be suboptimal in terms of number of measurements required to solve the problem [23].* While posing and solving the tensor recovery problem in a non-convex form tends to circumvent this issue, it brings about difficulties in terms of theoretically characterizing behavior of the associated recovery algorithm. In the next section, we present our proposed method for solving (4.9), while theoretical characterization of the proposed approach is an important contribution of this chapter.

4.3.1 Our Contributions

In the following we summarize the main contributions of this chapter:

1. We propose a new tensor regression model in (4.6), where we impose the parameter tensor $\mathbf{B}$ is $\underline{r}$-rank and $\underline{s}$-sparse. We formulate parameter estimation as a non-convex problem in (4.9), and we propose Algorithm 2 to solve it. In contrast, prior

works that study simultaneous imposition of multiple structures for regression either (i) assume that the tensors satisfy certain cubic structures [27], or (ii) formulate a convex problem for estimating the parameter tensor [28], which can lead to sub-optimal sample complexity [23].

2. We provide theoretical analysis to show that Algorithm 2 provides an approximately correct solution within a given number of algorithm steps, under a certain Restricted Isometry Property assumption—related to $\underline{r}-$rank and $\underline{s}-$sparse tensors—on the linear map. This assumption is reminiscent of the Restricted Isometry Property in the literature for sparse regression [69, 74] and low rank tensor regression [22, 72].

The rest of this chapter is organized as follows. In Sec. 4.4, we present our proposed algorithm for learning the tensor regression model in (4.6); whereas, in Sec. 4.5, we provide mathematical guarantees for the algorithm, based on a certain Restricted Isometry Property assumption on the linear map. Finally, in Sec. 5.3, we report results of extensive numerical experiments on synthetic data, while concluding remarks are presented in Sec. 5.4.

4.4 Estimation of $\underline{r}$-Rank and $\underline{s}$-Sparse Regression Tensors

In this section, we present a method for estimation of the structured parameter tensor $\mathbf{B}$ in the regression model (4.6), given the linear map $\mathcal{X}$, response vector $\mathbf{y}$, and the assumption that $\mathbf{B}$ is $\underline{r}$-rank and $\underline{s}$-sparse. Our method is inspired by the various projected gradient descent-based methods in the literature, where such methods have been employed for recovery of sparse vectors [69], low-rank matrices [71], and more recently, low rank tensors [22, 72]. The method, termed *tensor projected gradient descent* (TPGD), is summarized in Algorithm 2. The TPGD method consists of two steps. First we perform gradient descent iteration over the objective function in (4.9) (Step 4, Algorithm 2), and then, we project the iterate onto set $\mathcal{C}$, which is the set of $\underline{r}$-rank and $\underline{s}$-sparse tensors (Step 5, Algorithm 2). The projection operator, $\mathcal{H} : \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d} \rightarrow$

$\mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d}$, in Step 5 of Algorithm 2 is defined as:

$$\mathcal{H}(\widetilde{\mathbf{B}}) := \arg\min_{\widehat{\mathbf{B}} \in \mathcal{C}} \|\widetilde{\mathbf{B}} - \widehat{\mathbf{B}}\|_F^2. \tag{4.10}$$

Algorithm 2: Tensor Projected Gradient Descent (TPGD)

1: **Input:** Linear map $\mathcal{X}$, response vector $\mathbf{y}$, step size μ, sparsity tuple $\underline{s}$, rank tuple $\underline{r}$
2: **Initialize:** Tensor $\mathbf{B}^0$ and $k \leftarrow 0$
3: **while** Stopping criterion **do**
4: $\quad \widetilde{\mathbf{B}}^k \leftarrow \mathbf{B}^k - \mu \mathcal{X}^*(\mathcal{X}(\mathbf{B}^k) - \mathbf{y})$
5: $\quad \mathbf{B}^{k+1} \leftarrow \mathcal{H}(\widetilde{\mathbf{B}}^k)$
6: $\quad k \leftarrow k + 1$
7: **end while**
8: **return** Tensor $\mathbf{B}^* = \mathbf{B}^k$

In general, computation of tensor projections, such as the one given in (4.10), is considered to be an NP-hard problem [75]. Despite that, several algorithms have been proposed in the literature for computing low-rank tensor approximations corresponding to various notions of tensor decompositions [16, 76–78]. Although these approximation algorithms do not come with mathematical guarantees regarding the accuracy of tensor approximation, they have been employed successfuly in practice for parameter estimation in various examples of tensor regression models [22, 27, 72, 73, 78]. Correspondingly, since these approximation methods are not guaranteed to obtain the best tensor approximation, mathematical guarantees for the various parameter estimation methods *assume* the goodness of the tensor approximation step.

In a similar vein, in our mathematical guarantees for Algorithm 2 (Sec. 5.2), we *assume* that the projection step in (4.10) can be exactly computed. However, in our numerical simulations (Sec. 5.3), we employ Algorithm 3 for computation of the step in (4.10), where Algorithm 3 is essentially the Sparse Higher-Order SVD method [76], within which we employ the inverse power method from [79] for computation of the factor matrices. Despite the lack of mathematical guarantees for Algorithm 3, our numerical simulations show it can be effectively employed with Algorithm 2 to efficiently learn the regression model in (4.6) under certain conditions.

Algorithm 3: Sparse Higher-Order SVD

Input: Tensor $\widetilde{\mathbf{B}}$, sparsity tuple $\underline{s}$, rank tuple $\underline{r}$
for $j = 1, \ldots, d$ **do**
  $\overline{U}_j \leftarrow$ [column-wise arrangement of s_j-sparse principal components of $\widetilde{\mathbf{B}}_{(j)}$]
  $\overline{U}_j \leftarrow \overline{U}_j(:, 1 : r_j)$
end for
$\overline{\mathbf{S}} \leftarrow \widetilde{\mathbf{B}} \times_1 \overline{U}_1 \times_2 \overline{U}_2 \times_3 \cdots \times_d \overline{U}_d$
return Tensor $\overline{\mathbf{B}} = \overline{\mathbf{S}} \times_1 \overline{U}_1 \times_2 \overline{U}_2 \times_3 \cdots \times_d \overline{U}_d$

4.5 Convergence Analysis of Tensor Projected Gradient Descent

In this section we provide theoretical guarantees for TPGD (Algorithm 2), which, as explained earlier, is a projected gradient method to solve (4.9). Variants of the projected gradient method have been analyzed for recovery of sparse vectors [69], low-rank matrices [71], and low-rank tensors [22, 72] under the assumption that the linear map/measurement operator satisfies some variant of the restricted isometry property (RIP) [54]. Since different tensor decompositions induce different notions of tensor rank [22, 27], and different regression models lead to different measurement operators [22, 72], various notions of RIP have also been posed for various tensor decompositions and regression models. Before we present the notion of RIP assumed on the linear map in this work, let us define a set of $\underline{r}$-rank and $\underline{s}$-sparse tensors, with an additional constraint on the ℓ_1 norm of the associated core tensor:

$$\begin{aligned}\mathcal{G}_{\underline{r},\underline{s},\tau} =& \{\mathbf{S} \times_1 U_1 \times_2 U_2 \times_3 \cdots \times_d U_d : \mathbf{S} \in \mathbb{R}^{r_1 \times r_2 \times \cdots \times r_d}, \|\mathbf{S}\|_1 \leq \tau, \text{ and} \\ & U_i \in \mathbb{R}^{n_i \times r_i}, \|U_i(:, j)\|_0 \leq s_i, i \in [[d]], j \in [[r_i]]\}. \end{aligned} \tag{4.11}$$

For the recovery of $\underline{r}$-rank and $\underline{s}$-sparse tensors considered in this work, we consider the following notion of RIP on the linear map $\mathcal{X}$, which is followed by our first main theoretical result that characterizes the convergence behavior of TPGD under the assumption of an exact projection step (Step 5, Algorithm 2).

Definition 4.2 ($(\underline{r}, \underline{s}, \tau, \delta_{\underline{r},\underline{s},\tau})$-Restricted Isometry Property)**.** The restricted isometry constant $\delta_{\underline{r},\underline{s},\tau} \in (0, 1)$ of a linear map $\mathcal{X} : \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d} \to \mathbb{R}^m$ acting on tensors of

order d is the smallest quantity such that

$$(1-\delta_{\underline{r},\underline{s},\tau})\|\mathbf{Z}\|_F^2 \leq \|\mathcal{X}(\mathbf{Z})\|_2^2 \leq (1+\delta_{\underline{r},\underline{s},\tau})\|\mathbf{Z}\|_F^2 \tag{4.12}$$

for all tensors $\mathbf{Z} \in \mathcal{G}_{\underline{r},\underline{s},\tau}$.

Theorem 4.1 (Convergence of TPGD)**.** *Let* $\mathbf{y} = \mathcal{X}(\mathbf{B}) + \boldsymbol{\eta}$*, and let* $\mathbf{B}^0$ *be the tensor initialization in Algorithm 2. For some fixed* $\gamma \in (0,1)$*, if* $\mathcal{X} : \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d} \to \mathbb{R}^m$ *satisfies RIP in Definition 4.2 with* $\delta_{2\underline{r},\underline{s},2\tau} < \frac{\gamma}{4+\gamma}$*, then for* $b = \frac{1+3\delta_{2\underline{r},\underline{s},2\tau}}{1-\delta_{2\underline{r},\underline{s},2\tau}}$*, choosing some* $c_1 > 0$*, and fixing the step size* $\mu = \frac{1}{1+\delta_{2\underline{r},\underline{s},2\tau}}$*, the TPGD algorithm (Algorithm 2) obtains a solution* $\mathbf{B}^*$ *such that*

$$\|\mathbf{B}^* - \mathbf{B}\|_F^2 \leq \frac{2}{1-\delta_{2\underline{r},\underline{s},2\tau}}\Big(1 + c_1 + \frac{b}{1-\gamma}\Big)\|\boldsymbol{\eta}\|_2^2$$

in at most $\frac{1}{\log(\frac{1}{\gamma})}\log\Big(\frac{\|\mathbf{y}-\mathcal{X}(\mathbf{B}^0)\|_2^2}{c_1\|\boldsymbol{\eta}\|_2^2}\Big)$ *iterations.*

4.5.1 Discussion of Theorem 4.1

Let $c_0 := \frac{2}{1-\delta_{2\underline{r},\underline{s},2\tau}}\Big(1+c_1+\frac{b}{1-\gamma}\Big)\|\boldsymbol{\eta}\|_2^2$. Next, define the closed ball $\mathcal{B}(c_0, \mathbf{B})$ with center at $\mathbf{B}$ and radius c_0 as the set of all $\mathbf{Z} \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d}$ such that $\|\mathbf{Z}-\mathbf{B}\|_F^2 \leq d$. Further, let $t := \frac{1}{\log(\frac{1}{\gamma})}\log\Big(\frac{\|\mathbf{y}-\mathcal{X}(\mathbf{B}^0)\|_2^2}{c_1\|\boldsymbol{\eta}\|_2^2}\Big)$. With these definitions, it can be seen from Theorem 4.1 that the solution of TPGD will be in $\mathcal{B}(c_0, \mathbf{B})$ after t iterations of TPGD. Note that although the mathematical guarantees in this section depend on the $(\underline{r}, \underline{s}, \tau, \delta_{\underline{r},\underline{s},\tau})$-RIP property in Definition 4.2, we evaluate the property for a known family of linear maps in the next section.

Additionally, Theorem 4.1 also characterizes the impact of noise power $\|\boldsymbol{\eta}\|_2^2$ and RIP constant $\delta_{2\underline{r},\underline{s},2\tau}$ on convergence behavior of the TPGD algorithm. First, the radius of ball $\mathcal{B}(c_0, \mathbf{B})$ scales linearly with the noise power $\|\boldsymbol{\eta}\|_2^2$. Thus, the more the noise power, the less accurate may the solution of TPGD be and vice versa. Second, Theorem 4.1 shows that the smaller the RIP constant $\delta_{2\underline{r},\underline{s},2\tau}$, the smaller the radius of ball $\mathcal{B}(c_0, \mathbf{B})$. Thus, the larger the value of $\delta_{2\underline{r},\underline{s},2\tau}$, the less accurate may the solution of TPGD be and vice versa. Furthermore, observing that $\|\mathbf{y} - \mathcal{X}(\mathbf{B}^0)\|_2 \leq \sqrt{1+\delta_{2\underline{r},\underline{s},2\tau}}\|\mathbf{B} - \mathbf{B}^0\|_F + \|\boldsymbol{\eta}\|_2$ given $\mathbf{B}^0 \in \mathcal{B}(c_0, \mathbf{B})$, the theorem also shows that the rate of convergence of the algorithm is inversely related to the value of the RIP constant $\delta_{2\underline{r},\underline{s},2\tau}$ and the initialization distance $\|\mathbf{B} - \mathbf{B}^0\|_F$.

4.5.2 Proof of Theorem 4.1

Before we present a proof of Theorem 4.1, let us provide a lemma that involves analysis of any step that involves linear combination of $\underline{r}$-rank and $\underline{s}$-sparse tensors. A key step in proving Theorem 4.1 is to show that any linear combination of two $\underline{r}$-rank and $\underline{s}$-sparse tensors has rank at most $2\underline{r}$ and sparsity $\underline{s}$. We formally state this in the form of following lemma.

Lemma 4.1. *Let* $\mathbf{Z}_a \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d}$ *and* $\mathbf{Z}_b \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d}$ *be members of the set* $\mathcal{G}_{\underline{r},\underline{s},\tau}$*, where* $\underline{r} := (r_1, r_2, \ldots, r_d)$*,* $\underline{s} := (s_1, s_2, \ldots, s_d)$*, and* $\tau \in \mathbb{R}^+$*. Define* $\mathbf{Z}_c = \gamma_a \mathbf{Z}_a + \gamma_b \mathbf{Z}_c$*, where* $\gamma_a, \gamma_b \in \mathbb{R}$*. Then,* $\mathbf{Z}_c$ *is a member of the set* $\mathcal{G}_{2\underline{r},\underline{s},\kappa}$*, where* $\kappa = (|\gamma_a| + |\gamma_b|)\tau$*.*

The proof of this lemma is provided in Appendix 4.8.1. We are now ready to present a complete technical proof of Theorem 4.1.

Proof. Let $\mathcal{L}(\mathbf{Z}) := \|\mathbf{y} - \mathcal{X}(\mathbf{Z})\|_2^2$ be the loss function for any $\mathbf{Z} \in \mathbb{R}^{n_1 \times n_2 \times \ldots \times n_d}$. Then, we have

$$
\begin{aligned}
\mathcal{L}(\mathbf{B}^{k+1}) - \mathcal{L}(\mathbf{B}^k) &= \|\mathbf{y} - \mathcal{X}(\mathbf{B}^{k+1})\|_2^2 - \|\mathbf{y} - \mathcal{X}(\mathbf{B}^k)\|_2^2 \\
&= \|\mathcal{X}(\mathbf{B}^{k+1})\|_2^2 - \|\mathcal{X}(\mathbf{B}^k)\|_2^2 - 2\langle \mathbf{y}, \mathcal{X}(\mathbf{B}^{k+1} - \mathbf{B}^k)\rangle \\
&= \|\mathcal{X}(\mathbf{B}^{k+1})\|_2^2 + \|\mathcal{X}(\mathbf{B}^k)\|_2^2 - 2\|\mathcal{X}(\mathbf{B}^k)\|_2^2 - 2\langle \mathbf{y}, \mathcal{X}(\mathbf{B}^{k+1} - \mathbf{B}^k)\rangle \\
&= \|\mathcal{X}(\mathbf{B}^{k+1} - \mathbf{B}^k)\|_2^2 + 2\langle \mathcal{X}(\mathbf{B}^k), \mathcal{X}(\mathbf{B}^{k+1})\rangle - 2\langle \mathcal{X}(\mathbf{B}^k), \mathcal{X}(\mathbf{B}^k)\rangle \\
&\quad - 2\langle \mathbf{y}, \mathcal{X}(\mathbf{B}^{k+1} - \mathbf{B}^k)\rangle \\
&= \|\mathcal{X}(\mathbf{B}^{k+1} - \mathbf{B}^k)\|_2^2 + 2\langle \mathcal{X}(\mathbf{B}^k) - \mathbf{y}, \mathcal{X}(\mathbf{B}^{k+1} - \mathbf{B}^k)\rangle \\
&= \|\mathcal{X}(\mathbf{B}^{k+1} - \mathbf{B}^k)\|_2^2 + 2\langle \mathcal{A}^*(\mathcal{X}(\mathbf{B}^k) - \mathbf{y}), \mathbf{B}^{k+1} - \mathbf{B}^k\rangle \\
&\leq (1 + \delta_{2\underline{r},\underline{s},2\tau})\|\mathbf{B}^{k+1} - \mathbf{B}^k\|_F^2 + 2\langle \mathcal{A}^*(\mathcal{X}(\mathbf{B}^k) - \mathbf{y}), \mathbf{B}^{k+1} - \mathbf{B}^k\rangle,
\end{aligned}
\tag{4.13}
$$

where the last inequality follows from application of Definition 4.2 with Lemma 4.1.

For any $\mathbf{Z} \in \mathbb{R}^{n_1 \times n_2 \times \ldots \times n_d}$, define

$$\begin{aligned} g(\mathbf{Z}) &:= (1+\delta_{2\underline{r},\underline{s},2\tau})\|\mathbf{Z}-\mathbf{B}^k\|_F^2 + 2\langle \mathcal{A}^*(\mathcal{X}(\mathbf{B}^k)-\mathbf{y}), \mathbf{Z}-\mathbf{B}^k\rangle \\ &\overset{(a)}{=} (1+\delta_{2\underline{r},\underline{s},2\tau})\|\mathbf{Z}-\widetilde{\mathbf{B}}^k+\mu\mathcal{A}^*(\mathbf{y}-\mathcal{X}(\mathbf{B}^k))\|_F^2 \\ &\quad + 2\langle \mathcal{A}^*(\mathcal{X}(\mathbf{B}^k)-\mathbf{y}), \mathbf{Z}-\widetilde{\mathbf{B}}^k+\mu\mathcal{A}^*(\mathbf{y}-\mathcal{X}(\mathbf{B}^k))\rangle \\ &\overset{(b)}{=} (1+\delta_{2\underline{r},\underline{s},2\tau})\|\mathbf{Z}-\widetilde{\mathbf{B}}^k\|_F^2 - \frac{1}{1+\delta_{2\underline{r},\underline{s},2\tau}}\|\mathcal{A}^*(\mathbf{y}-\mathcal{X}(\mathbf{B}^k))\|_F^2, \end{aligned} \tag{4.14}$$

where (a) follows by substituting $\mathbf{B}^k = \widetilde{\mathbf{B}}^k + \mu\mathcal{A}^*(\mathcal{X}(\mathbf{B}^k)-\mathbf{y})$ and (b) follows by substituting $\mu = \frac{1}{1+\delta_{2\underline{r},\underline{s},2\tau}}$. Then, since $\|\mathbf{B}^{k+1}-\widetilde{\mathbf{B}}^k\|_F \le \|\mathbf{B}-\widetilde{\mathbf{B}}^k\|_F$, which follows from $\mathbf{B}^{k+1} = \mathcal{H}(\widetilde{\mathbf{B}}^k)$, we have $g(\mathbf{B}^{k+1}) \le g(\mathbf{B})$. Using $g(\mathbf{B}^{k+1}) \le g(\mathbf{B})$ with (4.13), we obtain

$$\begin{aligned} \mathcal{L}(\mathbf{B}^{k+1}) - \mathcal{L}(\mathbf{B}^k) &\le (1+\delta_{2\underline{r},\underline{s},2\tau})\|\mathbf{B}-\mathbf{B}^k\|_F^2 + 2\langle \mathcal{A}^*(\mathcal{X}(\mathbf{B}^k)-\mathbf{y}), \mathbf{B}-\mathbf{B}^k\rangle \\ &= 2\delta_{2\underline{r},\underline{s},2\tau}\|\mathbf{B}-\mathbf{B}^k\|_F^2 + (1-\delta_{2\underline{r},\underline{s},2\tau})\|\mathbf{B}-\mathbf{B}^k\|_F^2 \\ &\quad + 2\langle \mathcal{A}^*(\mathcal{X}(\mathbf{B}^k)-\mathbf{y}), \mathbf{B}-\mathbf{B}^k\rangle \\ &\le 2\delta_{2\underline{r},\underline{s},2\tau}\|\mathbf{B}-\mathbf{B}^k\|_F^2 + \|\mathcal{X}(\mathbf{B}-\mathbf{B}^k)\|_2^2 \\ &\quad + 2\langle \mathcal{A}^*(\mathcal{X}(\mathbf{B}^k)-\mathbf{y}), \mathbf{B}-\mathbf{B}^k\rangle \\ &= 2\delta_{2\underline{r},\underline{s},2\tau}\|\mathbf{B}-\mathbf{B}^k\|_F^2 + \|\mathcal{X}(\mathbf{B}-\mathbf{B}^k)\|_2^2 \\ &\quad + 2\langle \mathcal{X}(\mathbf{B}^k), \mathcal{X}(\mathbf{B}-\mathbf{B}^k)\rangle - 2\langle \mathbf{y}, \mathcal{X}(\mathbf{B}-\mathbf{B}^k)\rangle \\ &= 2\delta_{2\underline{r},\underline{s},2\tau}\|\mathbf{B}-\mathbf{B}^k\|_F^2 + \|\mathcal{X}(\mathbf{B})\|_2^2 - \|\mathcal{X}(\mathbf{B}^k)\|_2^2 - 2\langle \mathbf{y}, \mathcal{X}(\mathbf{B}-\mathbf{B}^k)\rangle \\ &= 2\delta_{2\underline{r},\underline{s},2\tau}\|\mathbf{B}-\mathbf{B}^k\|_F^2 + \|\mathbf{y}-\mathcal{X}(\mathbf{B})\|_2^2 - \|\mathbf{y}-\mathcal{X}(\mathbf{B}^k)\|_2^2 \\ &\le \frac{2\delta_{2\underline{r},\underline{s},2\tau}}{1-\delta_{2\underline{r},\underline{s},2\tau}}\|\mathcal{X}(\mathbf{B}-\mathbf{B}^k)\|_2^2 + \mathcal{L}(\mathbf{B}) - \mathcal{L}(\mathbf{B}^k), \end{aligned} \tag{4.15}$$

where the last two inequalities follow from application of Definition 4.2 with Lemma 4.1. Thus, we have

$$\mathcal{L}(\mathbf{B}^{k+1}) \le \frac{2\delta_{2\underline{r},\underline{s},2\tau}}{1-\delta_{2\underline{r},\underline{s},2\tau}}\|\mathcal{X}(\mathbf{B}-\mathbf{B}^k)\|_2^2 + \mathcal{L}(\mathbf{B}). \tag{4.16}$$

Using $\mathcal{X}(\mathbf{B}) = \mathbf{y}-\boldsymbol{\eta}$, we have

$$\|\mathcal{X}(\mathbf{B}-\mathbf{B}^k)\|_2^2 = \|\mathbf{y}-\mathcal{X}(\mathbf{B}^k)-\boldsymbol{\eta}\|_2^2 \le 2(\|\mathbf{y}-\mathcal{X}(\mathbf{B}^k)\|_2^2 + \|\boldsymbol{\eta}\|_2^2) = 2\big(\mathcal{L}(\mathbf{B}^k) + \|\boldsymbol{\eta}\|_2^2\big), \tag{4.17}$$

where the inequality follows since $(u+v)^2 \leq 2(u^2+v^2)$ for all $u, v \in \mathbb{R}$. Using (4.16) with (4.17), and observing $\mathcal{L}(\mathbf{B}) = \|\boldsymbol{\eta}\|_2^2$, we obtain

$$\begin{aligned}\mathcal{L}(\mathbf{B}^{k+1}) &\leq \frac{4\delta_{2\underline{r},\underline{s},2\tau}}{1-\delta_{2\underline{r},\underline{s},2\tau}}\Big(\mathcal{L}(\mathbf{B}^k) + \|\boldsymbol{\eta}\|_2^2\Big) + \|\boldsymbol{\eta}\|_2^2 \\ &= \frac{4\delta_{2\underline{r},\underline{s},2\tau}}{1-\delta_{2\underline{r},\underline{s},2\tau}}\mathcal{L}(\mathbf{B}^k) + \Big(1 + \frac{4\delta_{2\underline{r},\underline{s},2\tau}}{1-\delta_{2\underline{r},\underline{s},2\tau}}\Big)\|\boldsymbol{\eta}\|_2^2. \end{aligned} \tag{4.18}$$

Using $\delta_{2\underline{r},\underline{s},2\tau} < \frac{\gamma}{4+\gamma}$, $\gamma < 1$, and $b = \frac{1+3\delta_{2\underline{r},\underline{s},2\tau}}{1-\delta_{2\underline{r},\underline{s},2\tau}}$ yields

$$\mathcal{L}(\mathbf{B}^{k+1}) \leq \gamma\mathcal{L}(\mathbf{B}^k) + b\|\boldsymbol{\eta}\|_2^2. \tag{4.19}$$

Iterative application of this inequality leads to

$$\mathcal{L}(\mathbf{B}^k) \leq \gamma^k\mathcal{L}(\mathbf{B}^0) + \frac{b}{1-\gamma}\|\boldsymbol{\eta}\|_2^2 \tag{4.20}$$

for all $k \geq 1$.

Next, let us fix some $K \in \mathbb{Z}^+$. In order to obtain $\mathcal{L}(\mathbf{B}^K)$ that is small enough for some $c_1 > 0$, that is,

$$\mathcal{L}(\mathbf{B}^K) \leq \gamma^K\mathcal{L}(\mathbf{B}^0) + \frac{b}{1-\gamma}\|\boldsymbol{\eta}\|_2^2 \leq \Big(c_1 + \frac{b}{1-\gamma}\Big)\|\boldsymbol{\eta}\|_2^2, \tag{4.21}$$

the algorithm requires $K \geq \frac{1}{\log(\frac{1}{\gamma})}\log(\frac{\mathcal{L}(\mathbf{B}^0)}{c_1\|\boldsymbol{\eta}\|_2^2})$. Finally, using Definition 4.2 with Lemma 4.1,

$$\begin{aligned}\|\mathbf{B}^K - \mathbf{B}\|_F^2 &\leq \frac{1}{1-\delta_{2\underline{r},\underline{s},2\tau}}\|\mathcal{X}(\mathbf{B}^K - \mathbf{B})\|_2^2 \\ &\overset{\text{(c)}}{\leq} \frac{2}{1-\delta_{2\underline{r},\underline{s},2\tau}}\Big(\mathcal{L}(\mathbf{B}^K) + \|\boldsymbol{\eta}\|_2^2\Big) \\ &\overset{\text{(d)}}{\leq} \frac{2}{1-\delta_{2\underline{r},\underline{s},2\tau}}\Big(1 + c_1 + \frac{b}{1-\gamma}\Big)\|\boldsymbol{\eta}\|_2^2, \end{aligned} \tag{4.22}$$

where (c) and (d) follow from (4.17) and (4.21), respectively. □

4.6 Experimental Results

In this section, we analyze the performance of our proposed TPGD method (Algorithm 2) for learning tensor regression models, using numerical experiments on synthetic data. We compare the TPGD method with two tensor variants of the projected gradient

descent method, in which we consider low Tucker rank [22] and low CP rank [80] on the parameter tensor, respectively. Some relevant implementation details for these learning methods are as follows. For computation of $\mathcal{H}$ in Algorithm 2, we employ Algorithm 3 (Sparse Higher-Order SVD method [76]), within which we employ the inverse power method from [79] for computation of the factor matrices. For computation of the projection steps in the Tucker rank and the CP rank based methods, we employ the tensor toolboxes in [81] and [82], respectively.

For the synthetic-data experiments, we generate the $\underline{r}$-rank and $\underline{s}$-sparse tensor $\mathbf{B} \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d}$ in (4.6) as follows. We set $d = 3$, $n_1 = 50$, $n_2 = 50$, $n_3 = 20$, and in (4.7), we set $s_1 = 2$, $s_2 = 2$, $s_3 = 2$, and $r = 3$. For each $j \in [[d]]$, we generate the r column vectors $U_j(:, i)$, for all $i \in [[r]]$, such that $\|U_j(:, i)\|_0 \leq \underline{s}_j$. The $\underline{s}_j$ non-zero entries in $U_j(:, i)$ are chosen uniformly at random from $[[n_j]]$. Setting $a = 0.2$, we sample the non-zero entries in $U_j(:, i)$ from $(-1)^u(a + |z|)$, where u was drawn from a Bernoulli distribution with parameter 0.5 and z was drawn from a standard Gaussian distribution, Gaussian$(0, 1)$. To finally generate the parameter tensor $\mathbf{B}$, the entries of the core tensor $\mathbf{S}$ are sampled from a uniform distribution with parameters 0 and 1, and the tensor $\mathbf{B}$ is generated as in (4.7). To generate the response vector $\mathbf{y}$, the tensors $\{\mathbf{X}_i\}_{i=1}^m$ are generated such their entries are i.i.d. Gaussian$(0, 1/m)$, the noise vector $\boldsymbol{\eta}$ is sampled from Gaussian$(0, \sigma_z^2 I)$, and then the response vector $\mathbf{y}$ is generated as in (4.6).

We perform the experiments for noise variance $\sigma_z = 0.1$ as well as $\sigma_z = 0.4$, for various values of m. For each value of σ_z and m, (i) the parameter tensor $\mathbf{B}$, the linear map $\mathcal{X}$, and the response vector $\mathbf{y}$ are generated as explained in the previous paragraph, and (ii) a parameter estimate $\hat{\mathbf{B}}$ is computed using each of the learning methods. The performance of each learning method is characterized using the recovery error, which is defined as $\frac{\|\mathbf{B}-\hat{\mathbf{B}}\|_F}{\|\mathbf{B}\|_F}$. For each value of σ_z and m, the experiment is repeated 50 times, and the average recovery error is reported in Fig. 4.1. For all learning methods, algorithm parameters like $\underline{r}$, $\underline{s}$, Tucker rank, and CP rank are chosen in separate validation experiments, where the parameters are chosen such as to minimize the recovery error on validation datasets.

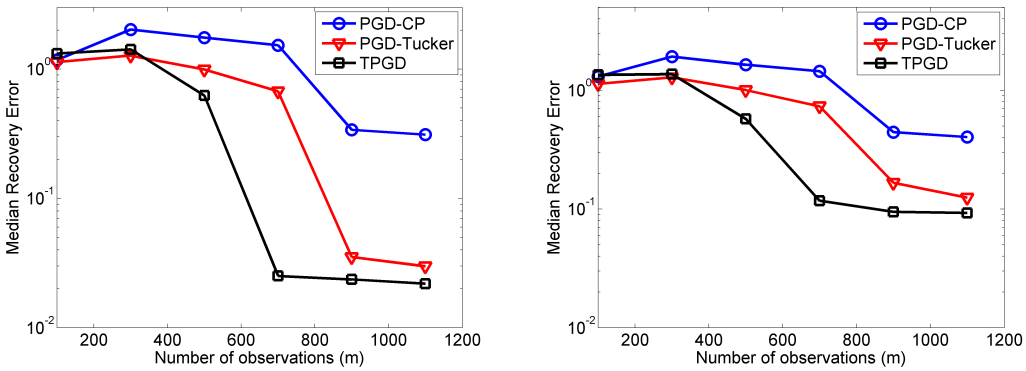
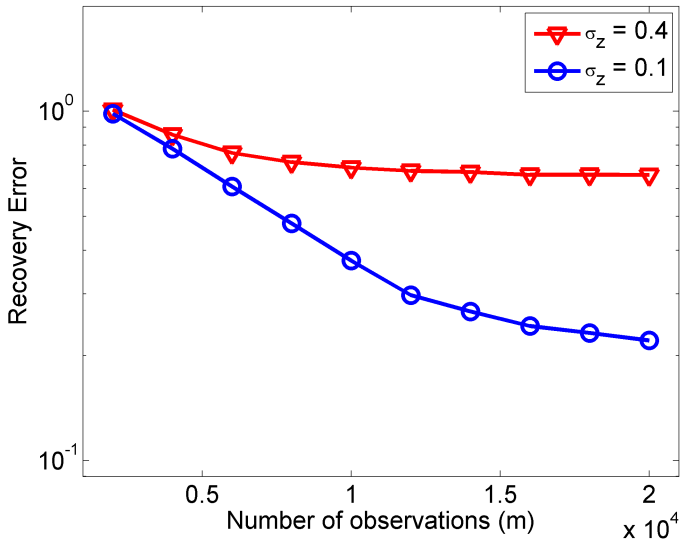


Figure 4.1: Comparison of TPGD with low Tucker rank and low CP rank based estimation over synthetic data. The mean error is reported over varying number of observations m, and the experiment is repeated for (a) $\sigma_z = 0.1$, (b) $\sigma_z = 0.4$. As σ_z decreases, the accuracy of the solution of TPGD increases. This is also reflected in Theorem 4.1: the lower the noise power, the more accurate the solution of TPGD.

Importantly, Fig. 4.1 shows the efficacy of the projection step $\mathcal{H}$ in Algorithm 2, since the proposed method demonstrates better sample complexity performance compared with the other learning methods. Despite the lack of theoretical guarantees for the projection step in Algorithm 2, the projection step can be computed accurately enough for us to be able to simultaneously exploit low rankness and sparsity in the parameter tensor. Furthermore, comparing Fig. 4.1a and Fig. 4.1b, we can see that as the noise power is reduced, the accuracy of the solution of TPGD increases, which is also what we learnt from our main result in Theorem 4.1.

Note that LASSO performs considerably worse than the other learning methods; thus, it's not included in Fig. 4.1 for clarity of plots. However, we plot the results for the LASSO method separately, in Fig. 4.2, for large sample sizes. Specifically, we perform the experiment for values of m ranging from 1000 to 20000, repeating the experiment for $\sigma_z = 0.1$ as well as $\sigma_z = 0.4$. As can be seen from Fig. 4.2, the recovery error for the LASSO method does decrease as we increase the sample size significantly. However, the LASSO method requires a much larger sample size to achieve recovery error that is comparable to the recovery error achieved by the other estimation methods in Fig. 4.1.



(a)

Figure 4.2: Recovery performance of LASSO method over synthetic data for large sample sizes. The recovery error is reported for values of m ranging from 1000 to 20000, and the experiment is repeated for (a) $\sigma_z = 0.1$, (b) $\sigma_z = 0.4$.

4.7 Conclusion

In this chapter, we proposed a new regression model that considers the simultaneous imposition of multiple structures on the parameter tensor, massively reducing the degrees of freedom in the model. We proposed an algorithm for parameter estimation, and we showed that the algorithm provides an approximately correct solution to the posed model under certain assumptions. Importantly, the simultaneous imposition of structures on the parameter tensor allowed us to provide better sample complexity bounds for parameter estimation using the proposed method, and in our experiments, we demonstrated the application of the proposed model and method in high-dimensional neuroimaging data analysis.

4.8 Appendix

4.8.1 Proof of Lemma 4.1

Since $\mathbf{Z}_a \in \mathcal{G}_{\underline{r},\underline{s},\tau}$, it can be expressed as

$$\mathbf{Z}_a = \mathbf{S}_a \times_1 U_{a,1} \times_2 U_{a,2} \cdots \times_d U_{a,d},$$

where $\mathbf{S}_a \in \mathbb{R}^{r_1 \times r_2 \times \cdots \times r_d}$ such that $\|\mathbf{S}_a\|_1 \leq \tau$, and $U_{a,i} \in \mathbb{R}^{n_i \times r_i}$, with $\|U_{a,i}(:,j)\|_0 \leq s_i$, $\forall i \in [[d]]$, $j \in [[r_i]]$. Similarly, since $\mathbf{Z}_b \in \mathcal{G}_{\underline{r},\underline{s},\tau}$, it can be expressed as

$$\mathbf{Z}_b = \mathbf{S}_b \times_1 U_{b,1} \times_2 U_{b,2} \cdots \times_d U_{b,d},$$

where $\mathbf{S}_b \in \mathbb{R}^{r_1 \times r_2 \times \cdots \times r_d}$ such that $\|\mathbf{S}_b\|_1 \leq \tau$, and $U_{b,i} \in \mathbb{R}^{n_i \times r_i}$, with $\|U_{b,i}(:,j)\|_0 \leq s_i$, $\forall i \in [[d]]$, $j \in [[r_i]]$. Let $\mathbf{Z}_c = \gamma_a \mathbf{Z}_a + \gamma_b \mathbf{Z}_c$, where $\gamma_a \in \mathbb{R}$, $\gamma_b \in \mathbb{R}$, so that $\mathbf{Z}_c$ is some linear combination of $\mathbf{Z}_a$ and $\mathbf{Z}_b$. Define the Cartesian product $\mathcal{D}_P := [[r_1]] \times [[r_2]] \times \cdots \times [[r_d]]$. Using the definition of $\mathcal{D}_P$, define $\mathbf{S}_c \in \mathbb{R}^{2r_1 \times 2r_2 \times \ldots \times 2r_d}$ where

$$\mathbf{S}_c(i_1, i_2, \ldots, i_d) = \begin{cases} \gamma_a \, \mathbf{S}_a(i_1, i_2, \ldots, i_d) & : (i_1, i_2, \ldots, i_d) \in \mathcal{D}_P \\ \gamma_b \, \mathbf{S}_b(i_1, i_2, \ldots, i_d) & : (i_1 - r_1, i_2 - r_2, \ldots, i_d - r_d) \in \mathcal{D}_P \\ 0 & : \text{otherwise} \end{cases}$$

for $(i_1, i_2, \ldots, i_d) \in [[2r_1]] \times [[2r_2]] \times \cdots \times [[2r_d]]$. Note that $\|\mathbf{S}_c\|_1 = \|\gamma_a \mathbf{S}_a\|_1 + \|\gamma_b \mathbf{S}_b\|_1 \leq (|\gamma_a| + |\gamma_b|)\tau$. Furthermore, for $i \in [[d]]$, define $U_{c,i} \in \mathbb{R}^{n_i \times 2r_i}$ such that $U_{c,i} := [U_{a,i} \;\; U_{b,i}]$. Finally, with these definitions, $\mathbf{Z}_c$ can be expressed as

$$\mathbf{Z}_c = \mathbf{S}_c \times_1 U_{c,1} \times_2 U_{c,2} \cdots \times_d U_{c,d},$$

where $\mathbf{S}_c \in \mathbb{R}^{2r_1 \times 2r_2 \times \ldots \times 2r_d}$ such that $\|\mathbf{S}_c\|_1 \leq (|\gamma_a| + |\gamma_b|)\tau$, and $U_{c,i} \in \mathbb{R}^{n_i \times 2r_i}$ such that $\|U_{c,i}(:,j)\|_0 \leq s_i$, for all $i \in [[d]]$, $j \in [[2r_i]]$. Therefore, $\mathbf{Z}_c$ is a member of the set $\mathcal{G}_{2\underline{r},\underline{s},\kappa}$, where $\kappa = (|\gamma_a| + |\gamma_b|)\tau$. $\square$

Chapter 5

Sample Complexity of Tensor Regression

In the previous chapter, we considered a new linear regression model for the case when predictors are tensor-valued, and we proposed a non-convex projected gradient descent-based method for estimating the parameter tensor. An important contribution of the previous chapter was to characterize the convergence behavior of the proposed method, based on a certain Restricted Isometry Property on the linear map. In this chapter, we evaluate the posed Restricted Isometry Property for the case when the predictors draw values from a sub-Gaussian distribution, and in the process, we characterize the sample complexity bound for parameter estimation. Importantly, our sample complexity bound only has a polylogarithmic dependence on n, where $n := \max\big\{n_i : i \in \{1, 2, \ldots, d\}\big\}$. In contrast, such sample complexity requirements in prior works pose a linear dependence on n. Furthermore, our real data experiments on an fMRI imaging dataset demonstrate the efficacy of the proposed regression model for neuroimaging data analysis. Specifically, our proposed model and method exhibit better classification performance on the neuroimaging dataset, demonstrating their applicability in settings where $\prod_i n_i \ggg m$.

5.1 Contributions

In the following we summarize the main contributions of this chapter:

1. We evaluate the posed Restricted Isometry Property (RIP) (Definition 4.2) for sub-Gaussian linear maps, and we characterize the sample complexity of parameter estimation. Specifically, we show that Algorithm 2 requires $\mathcal{O}\Big(\big(\prod_{i=1}^{d} r_i + \sum_{i=1}^{d} s_i r_i\big)\big(\log(3nd)\big)^2\Big)$ samples for providing an approximately correct solution, where $n := \max\{n_i : i \in [[m]]\}$. Importantly, our sample complexity bound only

has a polylogarithmic dependence on n. In contrast, prior works in tensor regression pose a sample complexity requirement that is either linear or super-linear in n [21–23].

2. In our synthetic experiments, we demonstrate the efficacy of the proposed method (Algorithm 2) for learning the proposed regression model in (4.6). In our real data experiments, we demonstrate the utility of our proposed model and method in diagnosis of attention deficit hyperactivity disorder (ADHD) using fMRI images. Importantly, these experiments show that despite the imposition of low rank and sparse structure on the parameter tensor $\mathbf{B}$, which leads to a massive decrease in degrees of freedom of the model, our model is not restrictive and is useful for neuroimaging data analysis.

The rest of the chapter is organized as follows. In Sec. 5.2, we evaluate the posed Restricted Isometry Property for sub-Gaussian linear maps and provide sample complexity bounds. In Sec. 5.3, we report results of numerical experiments on real data, while concluding remarks are presented in Sec. 5.4.

5.2 Evaluating the Restricted Isometry Property for Sample Complexity Analysis

In the previous chapter, we provided theoretical guarantees for recovery of the parameter tensor $\mathbf{B}$ using the TPGD method, based on assumption of the Restricted Isometry Property (Definition 4.2). In this section, we provide examples of linear maps that satisfy this property. Specifically, we consider linear maps in (4.6), $\mathcal{X}$, that denote the collection of tensors in (1.1), $\{\mathbf{X}_i\}_{i=1}^m$, such that the entries of $\{\mathbf{X}_i\}_{i=1}^m$ are independently drawn from zero-mean, unit-variance sub-Gaussian distributions. We denote such linear maps as sub-Gaussian linear maps. Before we evaluate the condition in Definition 4.2 for these maps, let us recall the definition of a sub-Gaussian random variable.

Definition 5.1. A zero-mean random variable $\mathcal{Z}$ is said to follow a sub-Gaussian distribution $subG(\alpha)$ if there exists a sub-Gaussian parameter $\alpha > 0$ such that $\mathbb{E}[\exp(\lambda \mathcal{Z})] \leq \exp\left(\frac{\alpha^2\lambda^2}{2}\right)$ for all $\lambda \in \mathbb{R}$.

In words, a $subG(\alpha)$ random variable is one whose moment generating function is dominated by that of a Gaussian random variable. Some common examples of sub-Gaussian random variables include:

- $\mathcal{Z} \sim \mathcal{N}(0, \alpha^2) \;\Rightarrow\; \mathcal{Z} \sim subG(\alpha)$.
- $\mathcal{Z} \sim \text{unif}(-\alpha, \alpha) \;\Rightarrow\; \mathcal{Z} \sim subG(\alpha)$.
- $|\mathcal{Z}| \leq \alpha, \mathbb{E}[\mathcal{Z}] = 0 \;\Rightarrow\; \mathcal{Z} \sim subG(\alpha)$.
- $\mathcal{Z} \sim \begin{cases} \alpha, & \text{with prob. } \frac{1}{2}, \\ -\alpha, & \text{with prob. } \frac{1}{2}, \end{cases} \Rightarrow\; \mathcal{Z} \sim subG(\alpha)$.

Next, we evaluate the Restricted Isometry Property (Definition 4.2) for sub-Gaussian linear maps.

Theorem 5.1. *Let the entries of $\{\mathbf{X}_i\}_{i=1}^m$ be independently drawn from zero-mean, $\frac{1}{m}$-variance $subG(\alpha)$ distributions. Define $\bar{n} := \max\{n_i : i \in [[d]]\}$. Then, for any $\delta, \varepsilon \in (0,1)$, the linear map $\mathcal{X}$ satisfies $\delta_{\underline{r},\underline{s},\tau} \leq \delta$ with probability at least $1-\varepsilon$ as long as*

$$m \geq \delta^{-2} \max\left\{ K_1\, \tau^2 \left(\prod_{i=1}^{d} r_i + \sum_{i=1}^{d} s_i r_i \right) \Big(\log(3\bar{n}d) \Big)^2,\; K_2 \log(\varepsilon^{-1}) \right\},$$

where the constants K_1, $K_2 > 0$ depend on τ and α.

5.2.1 Discussion

We compare the result in Theorem 5.1 with sample complexity bounds in the literature for estimation of the parameter tensor $\mathbf{B}$ in (4.6). Theoretically, we can pose the estimation problem as (i) low Tucker-rank recovery problem [22], or (ii) sparse recovery problem [74]. Thus, in this section, we first compare the sample complexity bound in Theorem 5.1 with complexity bounds from low rank recovery and sparse recovery literature. For ease of comparison, define $\bar{r} := \max\{r_1, r_2, \ldots, r_d\}$ and $\bar{s} := \max\{s_1, s_2, \ldots, s_d\}$. With these definitions, the sample requirement in Theorem 5.1 can be posed as $\Omega\Big(\big(\bar{r}^d + \bar{s}\,\bar{r}\,d\big) \big(\log(3\,\bar{n}\,d) \big)^2 \Big)$. Next, let's compare this complexity result with complexity bounds in aforementioned prior works.

Low Tucker-Rank Recovery

Among the many works that study the problem of estimating $\mathbf{B}$ under the imposition of low Tucker rank on $\mathbf{B}$ [10, 21–23], the most tight sample complexity bound has been shown to be $\Omega\big((\bar{r}^d + \bar{n}\,\bar{r}\,d)\log(d)\big)$ [22]. If we apply this complexity bound for estimating the parameter tensor $\mathbf{B}$ in (4.6), the sample complexity requirement scales linearly with $\bar{n}$. In contrast, since we consider sparsity on columns of the factor matrices within Tucker decomposition of $\mathbf{B}$, our sample complexity bound has a linear dependence on $\bar{s}$ and only a polylogarithmic dependence on $\bar{n}$, where $\bar{s} \ll \bar{n}$.

Sparse Recovery

The regression model in (4.6), or equivalently the model in (1.1), can be vectorized such that the model can be expressed as $y_i = \langle \mathrm{vec}(\mathbf{X}_i), \mathrm{vec}(\mathbf{B})\rangle + \eta_i$, $i \in [[m]]$, and the problem of recovering $\mathbf{B}$ can be posed as a sparse recovery problem. It has been shown that if the entries of $\mathrm{vec}(\mathbf{X}_i)$, $i \in [[m]]$, draw values from a Gaussian distribution, $\mathrm{vec}(\mathbf{B})$ can be recovered using $\mathcal{O}(k\log(\bar{n}^d/k))$ samples [83], where k is the number of non-zero entries in $\mathrm{vec}(\mathbf{B})$. The number of non-zero entries in $\mathrm{vec}(\mathbf{B})$ are upper bounded by $(\bar{s}\,\bar{r})^d$, which leads to a worst-case sample complexity requirement of $\mathcal{O}(d\,(\bar{s}\,\bar{r})^d \log(\bar{n}/\bar{s}\,\bar{r}))$. Thus, the sparse signal recovery literature poses a worst-case sample complexity requirement that has linear dependence on $d\,(\bar{s}\,\bar{r})^d$. In contrast, since we consider the multi-dimensional structure within $\mathbf{B}$, our sample complexity requirement has linear dependence on $\bar{r}^d + \bar{s}\,\bar{r}\,d$ only.

Finally, note that the number of free parameters in the parameter tensor $\mathbf{B}$ are on the order of $\prod_i r_i + \sum_i r_i s_i \log n_i$, which can be more conveniently expressed as $\bar{r}^d + \bar{s}\,\bar{r}\,d\log\bar{n}$. Thus, the posed sample complexity requirement of $\Omega\Big(\big(\bar{r}^d + \bar{s}\,\bar{r}\,d\big)\big(\log(3\,\bar{n}\,d)\big)^2\Big)$ in Theorem 5.1 is order-optimal up to a polylogarithmic factor.

5.2.2 Outline of the Proof

The general idea of the proof of Theorem 5.1 is similar to that of [71, Theorem 4.2], [84, Theorem 2.3], [85, Theorem 4.1], and [22, Theorem 2], where the main analytic challenge

is to analyze the complexity of the set that is hypothesized to contain the regression parameters. In this work, the challenge translates into characterizing the complexity of the set $\mathcal{G}_{\underline{r},\underline{s},\tau}$, for which we employ the notion of ϵ-nets and covering numbers, defined as follows.

Definition 5.2 (ϵ-nets and covering numbers)**.** Let (V, h) be a metric space, and let $T \subset V$. The set $X \subset T$ is called an ϵ-net of T with respect to the metric h if for any $T_i \in T$, $\exists X_i \in X$ such that $d(X_i, T_i) \leq \epsilon$. The minimum cardinality of an ϵ-net of T (with respect to the metric d) is called the covering number of T with respect to the metric d and is denoted by $\Psi(T, d, \epsilon)$ in this paper.

Next, we provide an outline to the proof of Theorem 5.1. In the first step, we provide an upper bound on the covering number of $\mathcal{G}_{\underline{r},\underline{s},\tau}$ with respect to the Frobenius norm, which forms our main contribution. In the second step, we employ a deviation bound from prior works [22, 85] to complete the proof of this theorem. A formal proof of Theorem 5.1 follows in Appendix 5.8.

Bound on Covering Number of $\mathcal{G}_{\underline{r},\underline{s},\tau}$

The following result provides a bound on the covering number of $\mathcal{G}_{\underline{r},\underline{s},\tau}$ with respect to the Frobenius norm:

Lemma 5.1. *For tuples $\underline{r} := (r_1, r_2, \ldots, r_d)$, $\underline{s} := (s_1, s_2, \ldots, s_d)$, and for any $\tau > 0$, the covering number of*

$$\begin{aligned}\mathcal{G}_{\underline{r},\underline{s},\tau} =&\{\mathbf{S} \times_1 U_1 \times_2 U_2 \times_3 \cdots \times_d U_d : \mathbf{S} \in \mathbb{R}^{r_1 \times r_2 \times \cdots \times r_d}, \|\mathbf{S}\|_1 \leq \tau, \text{ and} \\ & U_i \in \mathbb{R}^{n_i \times r_i}, \|U(:,j)\|_2 \leq 1, \|U_i(:,j)\|_0 \leq s_i, i \in [[d]], j \in [[r_i]]\}\end{aligned}$$

with respect to the metric $h_{\mathcal{G}}$ satisfies

$$\Psi(\mathcal{G}_{\underline{r},\underline{s},\tau}, h_{\mathcal{G}}, \epsilon) \leq \Big(\frac{3\tau(d+1)}{\epsilon}\Big)^{\prod_{i=1}^{d} r_i} \Big(\frac{3\bar{n}\tau(d+1)}{\epsilon}\Big)^{\sum_{i=1}^{d} s_i r_i}, \epsilon \in (0,1),$$

where $\bar{n} := \max\{n_i : i \in [[m]]\}$ and $h_{\mathcal{G}}(\mathbf{G}^{(1)}, \mathbf{G}^{(2)}) = \|\mathbf{G}^{(1)} - \mathbf{G}^{(2)}\|_F$ for any $\mathbf{G}^{(1)}, \mathbf{G}^{(2)} \in \mathcal{G}_{\underline{r},\underline{s},\tau}$.

Let us provide an outline to the proof of Lemma 5.1, while a formal proof is provided in Appendix 5.7. Define Cartesian product of metric spaces $(\mathcal{D}_{\mathbf{S}}, h_{\mathbf{S}})$, $(\mathcal{D}_{U_1}, h_{U_1})$, $(\mathcal{D}_{U_2}, h_{U_2})$, $\ldots$, $(\mathcal{D}_{U_d}, h_{U_d})$, that is

$$\mathcal{D}_P := \mathcal{D}_{\mathbf{S}} \times \mathcal{D}_{U_1} \times \mathcal{D}_{U_2} \times \cdots \times \mathcal{D}_{U_d},$$

where $\mathcal{D}_{\mathbf{S}} := \{\mathbf{S} \in \mathbb{R}^{r_1 \times r_2 \times \cdots \times r_d} : \|\mathbf{S}\|_1 \leq \tau\}$, $h_{\mathbf{S}}(\mathbf{S}^{(1)}, \mathbf{S}^{(2)}) := \frac{1}{\tau}\|\mathbf{S}^{(1)} - \mathbf{S}^{(2)}\|_1$ for any $\mathbf{S}^{(1)}, \mathbf{S}^{(2)} \in \mathcal{D}_{\mathbf{S}}$, $\mathcal{D}_{U_i} := \{U \in \mathbb{R}^{n_i \times r_i} : \|U(:,j)\|_2 \leq 1, \|U(:,j)\|_0 \leq s_i, j \in [[r_i]]\}$, and $h_{U_i}(U_i^{(1)}, U_i^{(2)}) = \|U_i^{(1)} - U_i^{(2)}\|_{1,2}$ for any $U_i^{(1)}, U_i^{(2)} \in \mathcal{D}_{U_i}$, for all $i \in [[d]]$. Next, we compute a bound on the covering number of $\mathcal{D}_P$ with respect to the metric h_P defined as

$$h_P(P^{(1)}, P^{(2)}) = \max\big\{ \max_{i \in [[d]]} \{\|U_i^{(1)} - U_i^{(2)}\|_{1,2}\}, \frac{1}{\tau}\|\mathbf{S}^{(1)} - \mathbf{S}^{(2)}\|_1 \big\},$$

where $P^{(1)}, P^{(2)} \in \mathcal{D}_P$, $\mathbf{S}^{(1)}, \mathbf{S}^{(2)} \in \mathcal{D}_{\mathbf{S}}$, and $U_i^{(1)}, U_i^{(2)} \in \mathcal{D}_{U_i}$, $i \in [[d]]$. Specifically, using Lemma 5.5, a bound on $\Psi(\mathcal{D}_P, h_P, \epsilon)$ can be obtained as

$$\Psi(\mathcal{D}_P, h_P, \epsilon) \leq \Psi(\mathcal{D}_{\mathbf{S}}, h_{\mathbf{S}}, \epsilon) \prod_{i=1}^{d} \Psi(\mathcal{D}_{U_i}, h_{U_i}, \epsilon). \tag{5.1}$$

Thus, to compute an upper bound on $\Psi(\mathcal{D}_P, h_P, \epsilon)$, we need upper bounds on $\Psi(\mathcal{D}_{\mathbf{S}}, h_{\mathbf{S}}, \epsilon)$ and $\Psi(\mathcal{D}_{U_i}, h_{U_i}, \epsilon)$, respectively. To obtain a bound on $\Psi(\mathcal{D}_{\mathbf{S}}, h_{\mathbf{S}}, \epsilon)$, we employ the following lemma, which is proved in Appendix 5.5.

Lemma 5.2. *Define $\mathcal{D}_{\mathbf{S}} := \{\mathbf{S} \in \mathbb{R}^{r_1 \times r_2 \times \cdots \times r_d} : \|\mathbf{S}\|_1 \leq \tau\}$ with distance measure $\|\,.\,\|_1$. Then the covering number of $\mathcal{D}_{\mathbf{S}}$ (with respect to the norm $\|\,.\,\|_1$) satisfies the bound*

$$\Psi(\mathcal{D}_{\mathbf{S}}, \|\,.\,\|_1, \epsilon) \leq \Big(\frac{3\,\tau}{\epsilon}\Big)^{\prod_{i=1}^{d} r_i}, \epsilon \in (0,1).$$

Similarly, to obtain a bound on $\Psi(\mathcal{D}_{U_i}, h_{U_i}, \epsilon)$ for any $i \in [[d]]$, we employ the following lemma, which is proved in Appendix 5.6.

Lemma 5.3. *Define $\mathcal{D}_U := \{U \in \mathbb{R}^{n \times r} : \|U(:,j)\|_2 \leq 1, \|U(:,j)\|_0 \leq s$ for all $j \in [[r]]\}$ with distance measure h_U, where $h_U(U^{(1)}, U^{(2)}) = \|U^{(1)} - U^{(2)}\|_{1,2}$ for any $U^{(1)}, U^{(2)} \in \mathcal{D}_U$. Then the covering number of $\mathcal{D}_U$ with respect to the metric h_U satisfies the bound*

$$\Psi(\mathcal{D}_U, h_U, \epsilon) \leq \Big(\frac{3n}{\epsilon}\Big)^{sr}, \epsilon \in (0,1).$$

Therefore, the bound in (5.1) can be evaluated using Lemma 5.2 and Lemma 5.3.

To finally derive a bound on the covering number of $\mathcal{G}_{\underline{r},\underline{s},\tau}$ with respect to the metric $h_{\mathcal{G}}$, define a mapping Φ such that

$$\Phi(\mathbf{S}, U_1, U_2, \ldots, U_d) = \mathbf{S} \times_1 U_1 \times_2 U_2 \times_3 \cdots \times_d U_d$$

where $(\mathbf{S}, U_1, U_2, \ldots, U_d) \in \mathcal{D}_P$. From this definition, it follows that $\Phi : \mathcal{D}_P \to \mathcal{G}_{\underline{r},\underline{s},\tau}$. We evaluate a constant $L \in \mathbb{R}$ such that $h_{\mathcal{G}}(\Phi(P^{(1)}), \Phi(P^{(2)})) \leq\ L\, h_P(P^{(1)}, P^{(2)})$, and then we employ Lemma 5.6 with (5.1) to obtain an upper bound on $\Psi(\mathcal{G}_{\underline{r},\underline{s},\tau}, h_{\mathcal{G}}, \epsilon)$.

Deviation Bound

Next, since $\delta_{\underline{r},\underline{s},\tau} = \sup_{\mathbf{Z} \in \mathcal{G}_{\underline{r},\underline{s},\tau}} \Big| \|\mathcal{X}(\mathbf{Z})\|_2^2 - \mathbb{E}\big[\|\mathcal{X}(\mathbf{Z})\|_2^2\big] \Big|$, we derive a probabilistic bound on the right hand side of this equality, using techniques similar to those in [22, 85]. Specifically, define ξ to be a random vector in $\mathbb{R}^{n_1\, n_2 \cdots n_d\, m}$ with independent entries from zero-mean, unit-variance, *subG*(B) random variables. Further, let $\mathbf{Z} \in \mathcal{G}_{\underline{r},\underline{s},\tau}$, and define $V_{\mathbf{Z}}$ to be a matrix in $\mathbb{R}^{m \times n_1\, n_2 \cdots n_d\, m}$ such that

$$V_{\mathbf{Z}} = \frac{1}{\sqrt{m}} \begin{bmatrix} \mathbf{z}^\top & 0 & 0 & \ldots & 0 \\ 0 & \mathbf{z}^\top & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \ldots & \mathbf{z}^\top \end{bmatrix},$$

where $\mathbf{z} \in \mathbb{R}^{n_1\, n_2 \cdots n_d \times 1}$ is the vectorized version of $\mathbf{Z}$. Then, we have the equivalence relationship $\mathcal{X}(\mathbf{Z}) = V_{\mathbf{Z}}\, \xi$. For ease of notation, let us further define a set $\mathcal{M} := \{V_{\mathbf{Z}} : \mathbf{Z} \in \mathcal{G}_{\underline{r},\underline{s},\tau}\}$. With this additional notation, we have $\delta_{\underline{r},\underline{s},\tau} = \sup_{M \in \mathcal{M}} \Big| \|M\xi\|_2^2 - \mathbb{E}\big[\|M\xi\|_2^2\big] \Big|$, and we apply the following theorem to obtain a deviation bound on the right hand side of this equality.

Theorem 5.2 ([22, 85])**.** *Let $\mathcal{M}_0$ be a set of matrices, and let ξ_0 be a random vector with independent entries from zero-mean, unit-variance, subG(α_0) random variables.*

For the set $\mathcal{M}_0$, define

$$d_F(\mathcal{M}_0) := \sup_{M\in\mathcal{M}_0} \|M\|_F,\ d_{2\to2}(\mathcal{M}_0) := \sup_{M\in\mathcal{M}_0} \|M\|_2,$$
$$\text{and } d_4(\mathcal{M}_0) := \sup_{M\in\mathcal{M}_0} \|M\|_{S_4} = \sup_{M\in\mathcal{M}_0} \Big(tr\big[(M^\top M)^2\big]\Big)^{\frac{1}{4}}.$$

Furthermore, let $\gamma_2(\mathcal{M}_0, \|\,.\,\|_2)$ be the Talagrand's γ_2-functional [86]. Finally, set

$$E = \gamma_2(\mathcal{M}_0, \|\,.\,\|_2)(\gamma_2(\mathcal{M}_0, \|\,.\,\|_2) + d_F(\mathcal{M}_0)) + d_F(\mathcal{M}_0)d_{2\to2}(\mathcal{M}_0)$$
$$V = d_4^2(\mathcal{M}_0), \text{ and } U = d_{2\to2}^2(\mathcal{M}_0).$$

Then, for $t > 0$,

$$\mathbb{P}\bigg(\sup_{M\in\mathcal{M}_0} \Big|\|M\xi\|_2^2 - \mathbb{E}\big[\|M\xi\|_2^2\big]\Big| \geq c_1E + t\bigg) \leq 2\exp\bigg(-c_2\min\bigg\{\frac{t^2}{V^2}, \frac{t}{U}\bigg\}\bigg),$$

where the constants c_1, c_2 depend on α_0.

For the application of Theorem 5.2, we evaluate bounds on the metrics $d_F(\mathcal{M})$, $d_{2\to2}(\mathcal{M})$, $d_4(\mathcal{M})$, and $\gamma_2(\mathcal{M}, \|\,.\,\|_2)$. However, the main analytical challenge in this application is evaluation of a bound on the Talagrand's γ_2-functional $\gamma_2(\mathcal{M}, \|\,.\,\|_2)$, which encompasses a geometric characterization of the metric space $(\mathcal{M}_0, \|\,.\,\|_2)$. We obtain a bound on the Talagrand's γ_2-functional using the following inequality [22, 86]:

$$\gamma_2(\mathcal{M}, \|\,.\,\|_2) \leq C\int_0^{d_{2\to2}(\mathcal{M})} \sqrt{\log\Psi(\mathcal{M}, \|\,.\,\|_2, \epsilon)}\, d\epsilon, \tag{5.2}$$

where $C > 0$ and $\Psi(\mathcal{M}, \|\,.\,\|_2, u)$ denotes the covering number of the metric space $(\mathcal{M},\ \|\,.\,\|_2)$ with respect to the metric $\|\,.\,\|_2$. Thus, we employ Theorem 5.2 with (5.2) and Lemma 5.1 to obtain a bound on $\sup_{M\in\mathcal{M}} \Big|\|M\xi\|_2^2 - \mathbb{E}\big[\|M\xi\|_2^2\big]\Big|$. A formal proof of Theorem 5.1 follows in Appendix 5.8.

5.3 Experimental Results

In this section, we perform experiments on real-world neuroimaging data to analyze the performance of the proposed TPGD method (Algorithm 2), which, as explained before, is a tensor variant of the projected gradient descent (PGD) method. As in the previous chapter, we compare TPGD with learning methods based on the imposition of sparsity,

low Tucker-rank and low CP-rank [16] on the parameter tensor $\mathbf{B}$, respectively. To analyze imposition of sparsity, we employ LASSO [13], and to analyze imposition of low Tucker-rank and low CP-rank, we employ Tucker-rank and CP-rank variants of the tensor projected gradient descent method, respectively.

To analyze the performance of TPGD for neuroimaging data analysis, we build a prediction model for attention deficit hyperactivity disorder (ADHD) diagnosis, using a preprocessed repository of ADHD-200 fMRI images [87] from the NYU Child Study Center. Specifically, we use preprocessed brain maps of fractional amplitude of low-frequency fluctuations (fALFF) [88] that were obtained using the Athena pipeline [24]. Note that fALFF is defined as the ratio of power within the low-frequency range (0.01-0.1 Hz) to that of the entire frequency range and as such it characterizes the intensity of spontaneous brain activity. Importantly, altered levels of fALFF have been reported in a sample of children with ADHD relative to controls [89], so fALFF brain maps form a useful feature space for predicting ADHD diagnosis.

The train data consists of fALFF brain maps of 227 individuals pertaining to NYU and NeuroImage. Each individual's fALFF map forms a third-order tensor $\mathbf{X}_i \in \mathbb{R}^{49\times 58\times 47}$, and the ADHD diagnosis y_i (1 = ADHD, 0 = normal control) forms the response, where $i \in [[m]]$. Thus, in our real data experiment, we have $m = 227$, $n_1 = 49$, $n_2 = 58$ and $n_3 = 47$. Given fALFF maps $\{\mathbf{X}_i\}_{i=1}^m$ and responses $\{y_i\}_{i=1}^m$, the task of learning the regression model in (1.1) is equivalent to learning the parameter tensor $\mathbf{B}$. We estimate the unknown parameter tensor using TPGD, PGD-Tucker, PGD-CP, and LASSO.

To analyze the performance of these learning methods, we employ separately provided test datasets for the NYU and the NeuroImage imaging sites, pertaining to fALFF maps of 41 subjects and 25 subjects, respectively. To analyze the performance for each method, we use the estimate of $\mathbf{B}$ to compute the responses for the test subjects using (1.1). If the computed response is more than 0.5 for a test subject, the subject is labeled with ADHD and vice versa. To evaluate the predictive power of each method using test data, we use the notion of (i) prediction accuracy, which is the ratio of subjects correctly labeled, (ii) sensitivity, which is the ratio of subjects diagnosed with ADHD

that are correctly labeled with ADHD, (*iii*) specificity, which is the ratio of subjects not diagnosed with ADHD that are correctly labeled as normal controls, and (*iv*) harmonic mean, which is the harmonic mean of sensitivity and specificity. The results of this experiment for NYU and NeuroImage imaging sites are shown in Table 5.1 and Table 5.2, respectively. Note that, in these experiments, the algorithmic parameters for each learning method were chosen in a five fold cross-validation experiment on the train data. Specifically, the algorithmic parameters were chosen such as to maximize the harmonic mean of sensitivity and specificity, as an average over the five folds of the cross-validation experiment.

Table 5.1: Comparison of TPGD with PGD-Tucker, PGD-CP, and LASSO for predicting ADHD diagnosis of 41 test subjects, who participated in the ADHD-200 Consortium at the New York University Child Study Center (NYU).

Method	Prediction accuracy	Sensitivity	Specificity	Harmonic mean
TPGD	0.561	0.517	0.667	**0.583**
PGD-Tucker	0.488	0.483	0.417	0.447
PGD-CP	0.439	0.448	0.417	0.432
LASSO	0.341	0.241	0.583	0.342

Table 5.2: Comparison of TPGD with PGD-Tucker, PGD-CP, and LASSO for predicting ADHD diagnosis of 25 test subjects, who participated in the ADHD-200 Consortium at The Donders Institute (NeuroImage).

Method	Prediction accuracy	Sensitivity	Specificity	Harmonic mean
TPGD	0.720	0.636	0.786	**0.703**
PGD-Tucker	0.640	0.727	0.500	0.593
PGD-CP	0.600	0.636	0.571	0.602
LASSO	0.600	0.455	0.714	0.556

5.4 Conclusion

In this chapter, we analyzed the sample complexity of learning the tensor-structured regression model proposed in the previous chapter. Specifically, we evaluated the restricted isometry property constant for the case of sub-Gaussian predictors, and in the process, we derived upper bound on the sample complexity of learning the regression

model under consideration. Finally, in our experiments with real-world data, we demonstrated that the posed tensor-structured regression model is not restrictive, and it can be effectively employed for neuroimaging data analysis.

5.5 Appendix

5.6 Proof of Lemma 5.3

The set $\mathcal{D}_U$ can be expressed as the Cartesian product of the sets $\mathcal{D}_U^{(j)} := \{x \in \mathbb{R}^n : \|x\|_0 \leq s, \|x\|_2 \leq 1\}$, $j \in [[r]]$. For any $j \in [[r]]$, since there are $\binom{n}{s}$ ways to choose the support of an s-sparse vector, we have

$$\Psi(\mathcal{D}_U^{(j)}, \|\,.\,\|_2, \epsilon) \leq \binom{n}{s}\left(\frac{3}{\epsilon}\right)^s, \tag{5.3}$$

with the application of Lemma 5.4. Then, the covering number of $\mathcal{D}_U$ with respect to the metric h_U, for any $\epsilon \in (0,1)$, satisfies the bound

$$\begin{aligned}\Psi(\mathcal{D}_U, h_U, \epsilon) &\overset{(a)}{\leq} \prod_{j=1}^{r} \Psi(\mathcal{D}_U^{(j)}, \|\,.\,\|_2, \epsilon) \overset{(b)}{\leq} \left[\binom{n}{s}\left(\frac{3}{\epsilon}\right)^s\right]^r \leq \frac{n^{sr}}{(s!)^r}\left(\frac{3}{\epsilon}\right)^{sr} = \left(\frac{3n}{(s!)^{\frac{1}{s}}\epsilon}\right)^{sr} \\ &\leq \left(\frac{3n}{\epsilon}\right)^{sr},\end{aligned}$$

where (a) and (b) follow from Lemma 5.5 and (5.3), respectively. □

5.7 Proof of Lemma 5.1

Recall the metric space $(\mathcal{D}_{\mathbf{S}}, h_{\mathbf{S}})$, where $\mathcal{D}_{\mathbf{S}} := \{\mathbf{S} \in \mathbb{R}^{r_1 \times r_2 \times \cdots \times r_d}\}$ and $h_{\mathbf{S}}(\mathbf{S}^{(1)}, \mathbf{S}^{(2)}) := \frac{1}{\tau}\|\mathbf{S}^{(1)} - \mathbf{S}^{(2)}\|_1$ for any $\mathbf{S}^{(1)}, \mathbf{S}^{(2)} \in \mathcal{D}_{\mathbf{S}}$. Using Lemma 5.2, the covering number of $\mathcal{D}_{\mathbf{S}}$ with respect to the metric $h_{\mathbf{S}}$ satisfies the bound

$$\Psi(\mathcal{D}_{\mathbf{S}}, h_{\mathbf{S}}, \epsilon) \leq \left(\frac{3}{\epsilon}\right)^{\prod_{i=1}^{d} r_i}, \epsilon \in (0,1).$$

Further, recall the metric space $(\mathcal{D}_{U_i}, h_{U_i})$, where $\mathcal{D}_{U_i} := \{U \in \mathbb{R}^{n_i \times r_i} : \|U(:,j)\|_2 \leq 1, \|U(:,j)\|_0 \leq s_i, j \in [[r_i]]\}$ and $h_{U_i}(U_i^{(1)}, U_i^{(2)}) := \|U_i^{(1)} - U_i^{(2)}\|_{1,2}$ for any $U_i^{(1)}, U_i^{(2)} \in \mathcal{D}_{U_i}$, $i \in [[d]]$. Using Lemma 5.3, the covering number of $\mathcal{D}_{U_i}$ with respect to the metric h_{U_i} satisfies the bound

$$\Psi(\mathcal{D}_{U_i}, h_{U_i}, \epsilon) \leq \left(\frac{3\bar{n}}{\epsilon}\right)^{s_i r_i}, \epsilon \in (0,1),$$

for any $i \in [[d]]$. Next, recall the metric space $(\mathcal{D}_P, h_P)$, where

$$\mathcal{D}_P := \mathcal{D}_{\mathbf{S}} \times \mathcal{D}_{U_1} \times \mathcal{D}_{U_2} \times \cdots \times \mathcal{D}_{U_d}, \text{ and}$$

$$h_P(P^{(1)}, P^{(2)}) = \max\Big\{ \max_{i \in [[d]]} \{\|U_i^{(1)} - U_i^{(2)}\|_{1,2}\}, \frac{1}{\tau}\|\mathbf{S}^{(1)} - \mathbf{S}^{(2)}\|_1 \Big\},$$

such that $P^{(1)}, P^{(2)} \in \mathcal{D}_P$, $\mathbf{S}^{(1)}, \mathbf{S}^{(2)} \in \mathcal{D}_{\mathbf{S}}$, and $U_i^{(1)}, U_i^{(2)} \in \mathcal{D}_{U_i}$, for any $i \in [[d]]$. Then, using Lemma 5.5, the covering number of $\mathcal{D}_P$ with respect to the h_P metric satisfies the bound

$$\Psi(\mathcal{D}_P, h_P, \epsilon) \leq \Big(\frac{3}{\epsilon}\Big)^{\prod_{i=1}^{d} r_i} \Big(\frac{3\bar{n}}{\epsilon}\Big)^{\sum_{i=1}^{d} s_i r_i}, \epsilon \in (0, 1). \tag{5.4}$$

To finally derive a bound on the covering number of $\mathcal{G}_{\underline{r},\underline{s},\tau}$, recall that we use the metric based on the Frobenius norm, denoted by $h_{\mathcal{G}}$, in order to cover the set $\mathcal{G}_{\underline{r},\underline{s},\tau}$. Further, recall the mapping Φ defined as

$$\Phi(\mathbf{S}, U_1, U_2, \ldots, U_d) = \mathbf{S} \times_1 U_1 \times_2 U_2 \times_3 \cdots \times_d U_d,$$

where $(\mathbf{S}, U_1, U_2, \ldots, U_d) \in \mathcal{D}_P$. From this definition, it follows that $\Phi : \mathcal{D}_P \to \mathcal{G}_{\underline{r},\underline{s},\tau}$. Then, given $P^{(1)}, P^{(2)} \in \mathcal{D}_P$, the claim is that

$$h_{\mathcal{G}}(\Phi(P^{(1)}), \Phi(P^{(2)})) \leq \tau(d+1) h_P(P^{(1)}, P^{(2)}), \tag{5.5}$$

which implies the mapping Φ is Lipschitz with a Lipschitz constant of $\tau\,(d+1)$. Using the claim in (5.5) with (5.4) and Lemma 5.6, the statement of this lemma follows. Next, we prove the claim in (5.5) to complete the proof of this lemma.

Let $\mathbf{G}_a, \mathbf{G}_b \in \mathcal{G}_{\underline{r},\underline{s},\tau}$ such that

$$\mathbf{G}_a = \mathbf{S}_a \times_1 U_{a,1} \times_2 U_{a,2} \times_3 \cdots \times_d U_{a,d} \text{ , and}$$
$$\mathbf{G}_b = \mathbf{S}_b \times_1 U_{b,1} \times_2 U_{b,2} \times_3 \cdots \times_d U_{b,d} \text{ ,}$$

where $\mathbf{S}_a, \mathbf{S}_b \in \mathcal{D}_{\mathbf{S}}$, and $U_{a,i}, U_{b,i} \in \mathcal{D}_{U_i}$, $i \in [[d]]$. Then, we have

$$
\begin{aligned}
h_{\mathcal{G}}(\mathbf{G}_a, \mathbf{G}_b) =& \|\mathbf{G}_a - \mathbf{G}_b\|_F \\
=& \|\mathbf{S}_a \times_1 U_{a,1} \times_2 U_{a,2} \times_3 \cdots \times_d U_{a,d} - \mathbf{S}_b \times_1 U_{b,1} \times_2 U_{b,2} \times_3 \cdots \times_d U_{b,d}\|_F \\
=& \|\mathbf{S}_a \times_1 U_{a,1} \times_2 U_{a,2} \times_3 \cdots \times_d U_{a,d} \\
& \pm \mathbf{S}_a \times_1 U_{a,1} \times_2 U_{a,2} \times_3 \cdots \times_{d-1} U_{a,d-1} \times_d U_{b,d} \\
& \pm \mathbf{S}_a \times_1 U_{a,1} \times_2 U_{a,2} \times_3 \cdots \times_{d-2} U_{a,d-2} \times_{d-1} U_{b,d-1} \times_d U_{b,d} \\
& \pm \cdots \pm \mathbf{S}_a \times_1 U_{b,1} \times_2 U_{b,2} \times_3 \cdots \times_{d-1} U_{b,d-1} \times_d U_{b,d} \\
& - \mathbf{S}_b \times_1 U_{b,1} \times_2 U_{b,2} \times_3 \cdots \times_{d-1} U_{b,d-1} \times_d U_{b,d}\|_F \\
\leq & \|\mathbf{S}_a \times_1 U_{a,1} \times_2 U_{a,2} \times_3 \cdots \times_{d-1} U_{a,d-1} \times_d (U_{a,d} - U_{b,d})\|_F \\
& + \|\mathbf{S}_a \times_1 U_{a,1} \times_2 U_{a,2} \times_3 \cdots \times_{d-2} U_{a,d-2} \times_{d-1} (U_{a,d-1} - U_{b,d-1}) \times_d U_{b,d}\|_F \\
& + \cdots + \|\mathbf{S}_a \times_1 (U_{a,1} - U_{b,1}) \times_2 U_{b,2} \times_3 \cdots \times_{d-1} U_{b,d-1} \times_d U_{b,d}\|_F \\
& + \|(\mathbf{S}_a - \mathbf{S}_b) \times_1 U_{b,1} \times_2 U_{b,2} \times_3 \cdots \times_{d-1} U_{b,d-1} \times_d U_{b,d}\|_F, \qquad (5.6)
\end{aligned}
$$

where $\pm\mathbf{V}$ denotes $+\mathbf{V} - \mathbf{V}$ for any tensor $\mathbf{V} \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d}$. Define $b_j := \|\mathbf{S}_a \times_1 U_{a,1} \times_2 \cdots \times_{j-1} U_{a,j-1} \times_j (U_{a,j} - U_{b,j}) \times_{j+1} U_{b,j+1} \times_{j+2} \cdots \times_d U_{b,d}\|_F$. With this definition, (5.6) can be re-written as

$$
h_{\mathcal{G}}(\mathbf{G}_a, \mathbf{G}_b) \leq \sum_{j=1}^{d} b_j + \|(\mathbf{S}_a - \mathbf{S}_b) \times_1 U_{b,1} \times_2 U_{b,2} \times_3 \cdots \times_{d-1} U_{b,d-1} \times_d U_{b,d}\|_F. \quad (5.7)
$$

We will bound the first d terms and the last term in (5.7) separately. Beginning with any term from among the first d terms in (5.7), for any $j \in [[d]]$, we have

$$
\begin{aligned}
b_j^2 &= \|\mathbf{S}_a \times_1 U_{a,1} \times_2 \cdots \times_{j-1} U_{a,j-1} \times_j (U_{a,j} - U_{b,j}) \times_{j+1} U_{b,j+1} \times_{j+2} \cdots \times_d U_{b,d}\|_F^2 \\
&= \sum_{i_1, i_2, \ldots, i_d} \big[\big(\mathbf{S}_a \times_1 U_{a,1} \times_2 \cdots \times_{j-1} U_{a,j-1} \times_j (U_{a,j} - U_{b,j}) \times_{j+1} U_{b,j+1} \times_{j+2} \cdots \\
&\quad \times_d U_{b,d} \big)(i_1, i_2, \ldots, i_d) \big]^2 \\
&= \sum_{i_1, i_2, \ldots, i_d} \Big(\sum_{k_1, k_2, \ldots, k_d} \mathbf{S}_a(k_1, k_2, \ldots, k_d)\, U_{a,1}(i_1, k_1)\, U_{a,2}(i_2, k_2) \cdots \\
&\quad \big(U_{a,j} - U_{b,j}\big)(i_j, k_j) \cdots U_{b,d}(i_d, k_d) \Big)
\end{aligned}
$$

$$\Big(\sum_{l_1,l_2,\ldots,l_d} \mathbf{S}_a(l_1,l_2,\ldots,l_d)\, U_{a,1}(i_1,l_1)\, U_{a,2}(i_2,l_2)\, \cdots\, \big(U_{a,j}-U_{b,j}\big)(i_j,l_j)\, \cdots\, U_{b,d}(i_d,l_d)\Big)$$

$$= \sum_{i_1,i_2,\ldots,i_d}\ \sum_{k_1,k_2,\ldots,k_d}\ \sum_{l_1,l_2,\ldots,l_d} \mathbf{S}_a(k_1,k_2,\ldots,k_d)\, \mathbf{S}_a(l_1,l_2,\ldots,l_d)\, U_{a,1}(i_1,k_1)\, U_{a,1}(i_1,l_1)$$

$$\cdots\, U_{a,2}(i_2,k_2)\, U_{a,2}(i_2,l_2)\, \cdots\, \big(U_{a,j}-U_{b,j}\big)(i_j,k_j)\, \big(U_{a,j}-U_{b,j}\big)(i_j,l_j)\, \cdots$$

$$U_{b,d}(i_d,k_d)\, U_{b,d}(i_d,l_d)$$

$$= \sum_{k_1,k_2,\ldots,k_d}\ \sum_{l_1,l_2,\ldots,l_d} \mathbf{S}_a(k_1,k_2,\ldots,k_d)\, \mathbf{S}_b(l_1,l_2,\ldots,l_d) \sum_{i_1} U_{a,1}(i_1,k_1)\, U_{a,1}(i_1,l_1)$$

$$\sum_{i_2} U_{a,2}(i_2,k_2)\, U_{a,2}(i_2,l_2)\, \cdots \sum_{i_j} \big(U_{a,j}-U_{b,j}\big)(i_j,k_j)\, \big(U_{a,j}-U_{b,j}\big)(i_j,l_j)\, \cdots$$

$$\sum_{i_d} U_{b,d}(i_d,k_d)\, U_{b,d}(i_d,l_d)$$

$$\overset{\text{(a)}}{\leq} \sum_{k_1,k_2,\ldots,k_d}\ \sum_{l_1,l_2,\ldots,l_d} \mathbf{S}_a(k_1,k_2,\ldots,k_d)\, \mathbf{S}_a(l_1,l_2,\ldots,l_d)$$

$$\sum_{i_j} \big(U_{a,j}-U_{b,j}\big)(i_j,k_j)\big(U_{a,j}-U_{b,j}\big)(i_j,l_j)$$

$$\leq\ \|U_{a,j}-U_{b,j}\|_{1,2}^2 \sum_{k_1,k_2,\ldots,k_d}\ \sum_{l_1,l_2,\ldots,l_d} \mathbf{S}_a(k_1,k_2,\ldots,k_d)\, \mathbf{S}_a(l_1,l_2,\ldots,l_d)$$

$$\leq\ \|U_{a,j}-U_{b,j}\|_{1,2}^2\, \|\mathbf{S}_a\|_1\, \|\mathbf{S}_a\|_1 \leq\ \|U_{a,j}-U_{b,j}\|_{1,2}^2\, \tau^2\,, \tag{5.8}$$

where (a) follows since $\mathbf{u}^\top \mathbf{v} \leq 1$ for any column vectors u and v such that $\|u\|_2 \leq 1$ and $\|v\|_2 \leq 1$. Similarly, to bound the last term in (5.7), note that

$$\|\big(\mathbf{S}_a-\mathbf{S}_b\big) \times_1 U_{b,1} \times_2 U_{b,2} \times_3 \cdots \times_{d-1} U_{b,d-1} \times_d U_{b,d}\|_F^2$$

$$= \sum_{i_1,i_2,\ldots,i_d}\ \sum_{k_1,k_2,\ldots,k_d} \big(\mathbf{S}_a-\mathbf{S}_b\big)(k_1,k_2,\ldots,k_d)\, U_{b,1}(i_1,k_1)\, U_{b,2}(i_2,k_2)\, \cdots\, U_{b,d}(i_d,k_d)$$

$$\sum_{l_1,l_2,\ldots,l_d} \big(\mathbf{S}_a-\mathbf{S}_b\big)(l_1,l_2,\ldots,l_d)\, U_{b,1}(i_1,l_1)\, U_{b,1}(i_2,l_2)\, \cdots\, U_{b,d}(i_d,l_d)$$

$$= \sum_{i_1,i_2,\ldots,i_d}\ \sum_{k_1,k_2,\ldots,k_d}\ \sum_{l_1,l_2,\ldots,l_d} \big(\mathbf{S}_a-\mathbf{S}_b\big)(k_1,k_2,\ldots,k_d)\, \big(\mathbf{S}_a-\mathbf{S}_b\big)(l_1,l_2,\ldots,l_d)$$

$$U_{b,1}(i_1,k_1)\, U_{b,1}(i_1,l_1)\, U_{b,2}(i_2,k_2)\, U_{b,2}(i_2,l_2)\, \cdots\, U_{b,d}(i_d,k_d)\, U_{b,d}(i_d,l_d)$$

$$= \sum_{k_1,k_2,\ldots,k_d}\ \sum_{l_1,l_2,\ldots,l_d} \big(\mathbf{S}_a-\mathbf{S}_b\big)(k_1,k_2,\ldots,k_d)\, \big(\mathbf{S}_a-\mathbf{S}_b\big)(l_1,l_2,\ldots,l_d) \tag{5.9}$$

$$\sum_{i_1} U_{b,1}(i_1,k_1)\, U_{b,1}(i_1,l_1)\, \sum_{i_2} U_{b,2}(i_2,k_2)\, U_{b,2}(i_2,l_2)\, \cdots \sum_{i_d} U_{b,d}(i_d,k_d)\, U_{b,d}(i_d,l_d)$$

$$\overset{\text{(b)}}{\leq} \sum_{k_1,k_2,\ldots,k_d}\ \sum_{l_1,l_2,\ldots,l_d} \big(\mathbf{S}_a - \mathbf{S}_b\big)(k_1,k_2,\ldots,k_d)\, \big(\mathbf{S}_a - \mathbf{S}_b\big)(l_1,l_2,\ldots,l_d)$$

$$\leq \|\mathbf{S}_a - \mathbf{S}_b\|_1^2, \tag{5.10}$$

where, again, (b) follows since $\mathbf{u}^\top \mathbf{v} \leq 1$ for any column vectors u and v such that $\|u\|_2 \leq 1$ and $\|v\|_2 \leq 1$. Finally, using (5.7) with (5.8) and (5.10), we obtain

$$\begin{aligned} h_{\mathcal{G}}(\mathbf{G}_a, \mathbf{G}_b) &\leq \sum_{j=1}^{d} \tau\, \|U_{a,j} - U_{b,j}\|_{1,2} + \|\mathbf{S}_a - \mathbf{S}_b\|_1 \\ &\leq (d+1)\,\tau\, \max\Big\{ \max_{j\in[[d]]} \{\|U_{a,j} - U_{b,j}\|_{1,2}\}, \frac{1}{\tau}\, \|\mathbf{S}_a - \mathbf{S}_b\|_1 \Big\} \\ &= (d+1)\,\tau\, h_P(P^{(1)}, P^{(2)}), \end{aligned} \tag{5.11}$$

which proves the claim in (5.5). □

5.8 Proof of Theorem 5.1

We employ Theorem 5.2 with Lemma 5.1 to obtain a probabilistic bound on the restricted isometry property constant in Definition 4.2. Before we can employ Theorem 5.2, we need to evaluate bounds on the quantities $d_F(\mathcal{M})$, $d_{2\to 2}(\mathcal{M})$, $d_4(\mathcal{M})$, and $\gamma_2(\mathcal{M}, \|\,.\,\|_2)$, which we obtain as follows. We obtain a bound on $d_F(\mathcal{M})$ as

$$d_F(\mathcal{M}) = \sup_{M\in\mathcal{M}} \|M\|_F \overset{\text{(a)}}{=} \sup_{\mathbf{Z}\in\mathcal{G}_{\underline{r},\underline{s},\tau}} \|\mathbf{Z}\|_F \overset{\text{(b)}}{\leq} \tau, \tag{5.12}$$

where (a) follows from the definition of $\mathcal{M}$ and (b) follows from the definition of $\mathcal{G}_{\underline{r},\underline{s},\tau}$. Next, to obtain a bound on $d_{2\to 2}(\mathcal{M})$ and $d_4(\mathcal{M})$, note that for any $\mathbf{Z} \in \mathcal{G}_{\underline{r},\underline{s},\tau}$ we have

$$V_{\mathbf{Z}} V_{\mathbf{Z}}^\top = \frac{1}{m} \begin{bmatrix} \mathbf{z}^\top\mathbf{z} & 0 & 0 & \ldots & 0 \\ 0 & \mathbf{z}^\top\mathbf{z} & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \ldots & \mathbf{z}^\top\mathbf{z} \end{bmatrix} = \frac{\|\mathbf{z}\|_2^2}{m}\, \mathbb{I}_{m\times m},$$

which leads to

$$d_{2\to 2}(\mathcal{M}) = \sup_{M\in\mathcal{M}} \|M\|_{\|\,.\,\|_2} = \sup_{\mathbf{Z}\in\mathcal{G}_{\underline{r},\underline{s},\tau}} \frac{\|\mathbf{Z}\|_F}{\sqrt{m}} \leq \frac{\tau}{\sqrt{m}}, \text{ and} \tag{5.13}$$

$$d_4^4(\mathcal{M}) = \sup_{M\in\mathcal{M}} \|M\|_{S_4}^4 = \sup_{M\in\mathcal{M}} tr\Big[(M^\top M)^2\Big] = \sup_{M\in\mathcal{M}} tr\Big[(MM^\top)^2\Big]$$
$$= \sup_{\mathbf{Z}\in\mathcal{G}_{\underline{r},\underline{s},\tau}} tr\Big[\Big(\frac{\|\mathbf{Z}\|_F^2}{m}\,\mathbb{I}_{m\times m}\Big)^2\Big] = \sup_{\mathbf{Z}\in\mathcal{G}_{\underline{r},\underline{s},\tau}} \frac{\|\mathbf{Z}\|_F^4}{m^2}\, tr\Big[\mathbb{I}_{m\times m}\Big] \;\leq\; \frac{\tau^4}{m}. \quad (5.14)$$

Finally, to obtain a bound on the Talagrand's γ_2-functional, we use the following bound [22, 86]:

$$\gamma_2(\mathcal{M}, \|\,.\,\|_2) \leq C\int_0^{d_{2\to 2}(\mathcal{M})} \sqrt{\log \Psi(\mathcal{M}, \|\,.\,\|_2, u)}\; du, \quad (5.15)$$

where $C > 0$ and $\Psi(\mathcal{M}, \|\,.\,\|_2, u)$ denotes the covering number of the metric space $(\mathcal{M}, \|\,.\,\|_2)$ with respect to the metric $\|\,.\,\|_2$. We now employ (5.15) with Lemma 5.1 to obtain a bound on $\gamma_2(\mathcal{M}, \|\,.\,\|_2)$ as

$$\gamma_2(\mathcal{M}, \|\,.\,\|_2) \leq C\int_0^{d_{2\to 2}(\mathcal{M})} \sqrt{\log \Psi(\mathcal{M}, \|\,.\,\|_2, u)}\; du \leq C\int_0^{\frac{\tau}{\sqrt{m}}} \sqrt{\log \Psi(\mathcal{M}, \|\,.\,\|_2, u)}\; du$$
$$= \frac{C}{\sqrt{m}}\int_0^{\tau} \sqrt{\log \Psi(\mathcal{M}, \|.\|_F, \tilde{u})}\; d\tilde{u}$$
$$\overset{(c)}{\leq} \frac{C}{\sqrt{m}}\int_0^{\tau} \sqrt{\Big(\prod_{i=1}^d r_i + \sum_{i=1}^d s_i r_i\Big)\log\Big(\frac{3\bar{n}\tau(d+1)}{\tilde{u}}\Big)}\; d\tilde{u}$$
$$= C\sqrt{\frac{\prod_{i=1}^d r_i + \sum_{i=1}^d s_i r_i}{m}}\int_0^{\tau} \sqrt{\log\Big(\frac{3\bar{n}\tau(d+1)}{\tilde{u}}\Big)}\; d\tilde{u}$$
$$\overset{(d)}{\leq} C\sqrt{\frac{\prod_{i=1}^d r_i + \sum_{i=1}^d s_i r_i}{m}}\int_0^{\tau} \log\Big(\frac{3\bar{n}\tau(d+1)}{\tilde{u}}\Big)\; d\tilde{u}$$
$$= C\sqrt{\frac{\prod_{i=1}^d r_i + \sum_{i=1}^d s_i r_i}{m}}\;\Big[\tau\log\big(3\bar{n}(d+1)\big) + \tau\Big]$$
$$= C\sqrt{\frac{\tau^2\Big(\prod_{i=1}^d r_i + \sum_{i=1}^d s_i r_i\Big)}{m}}\;\Big[\log\big(3\bar{n}(d+1)\big) + 1\Big]$$
$$= \widetilde{C}\sqrt{\frac{\tau^2\Big(\prod_{i=1}^d r_i + \sum_{i=1}^d s_i r_i\Big)}{m}}\;\log\big(3\bar{n}d\big), \quad (5.16)$$

where $\widetilde{C} > 0$, (c) follows from Lemma 5.1, and (d) follows since $\sqrt{\log_b(x/a)} \leq \log_b(x/a)$ for $x/a \in \mathbb{R}^+$, $b \in \mathbb{R}^+$, $x \geq a\,b$. Now that we have evaluated bounds on $d_F(\mathcal{M})$,

$d_{2\to2}(\mathcal{M})$, $d_4(\mathcal{M})$, and $\gamma_2(\mathcal{M}, \|\,.\,\|_2)$, we can evaluate the quantities E, U, and V in Theorem 5.2. Evaluating a bound on E, we get

$$\begin{aligned}
E &= \gamma_2(\mathcal{M}, \|\,.\,\|_2)\Big(\gamma_2(\mathcal{M}, \|\,.\,\|_2) + d_F(\mathcal{M})\Big) + d_F(\mathcal{M}) d_{2\to2}(\mathcal{M}) \\
&= \gamma_2(\mathcal{M}, \|\,.\,\|_2)^2 + \gamma_2(\mathcal{M}, \|\,.\,\|_2) d_F(\mathcal{M}) + d_F(\mathcal{M}) d_{2\to2}(\mathcal{M}) \\
&\overset{\text{(e)}}{\leq} \widetilde{C}^2 \, \frac{\tau^2\Big(\prod\limits_{i=1}^{d} r_i + \sum\limits_{i=1}^{d} s_i r_i\Big)}{m} \Big(\log\big(3\bar{n}d\big)\Big)^2 + \widetilde{C}\,\tau \sqrt{\frac{\tau^2\Big(\prod\limits_{i=1}^{d} r_i + \sum\limits_{i=1}^{d} s_i r_i\Big)\, \log\big(3\bar{n}d\big)}{m}} + \frac{\tau^2}{\sqrt{m}} \\
&\overset{\text{(f)}}{\leq} \frac{\delta^2\,\widetilde{C}^2}{K_1} + \frac{\tau\,\delta\,\widetilde{C}}{\sqrt{K_1}} + \frac{\tau\,\delta}{\sqrt{K_1}} \overset{\text{(g)}}{\leq} \frac{\delta\,\widetilde{C}^2}{K_1} + \frac{\tau\,\delta\,\widetilde{C}}{\sqrt{K_1}} + \frac{\tau\,\delta}{\sqrt{K_1}} \;\leq\; \frac{\delta\,\big(\widetilde{C}^2 + \widetilde{C}\,\tau + \tau\big)}{\min\{K_1, \sqrt{K_1}\}}, \quad (5.17)
\end{aligned}$$

where (e) follows from application of (5.12) and (5.13) with (5.16), (f) follows from the bound on m, and (g) follows since $\delta \in (0,1)$. Setting $K_1 \geq \max\Big\{\Big(2c_1(\widetilde{C}^2 + \widetilde{C}\,\tau + \tau)\Big)^2, 2c_1(\widetilde{C}^2 + \widetilde{C}\,\tau + \tau)\Big\}$ in (5.17) for some $c_1 > 0$, we obtain

$$c_1 E \leq \frac{\delta\, c_1\,\big(\widetilde{C}^2 + \widetilde{C}\,\tau + \tau\big)}{\min\{K_1, \sqrt{K_1}\}} \leq \frac{\delta}{2}. \quad (5.18)$$

Next, we can evaluate bounds on U and V as

$$U = d_{2\to2}^2(\mathcal{M}) \overset{\text{(h)}}{\leq} \frac{\tau^2}{m}, \text{ and} \quad (5.19)$$

$$V = d_4^2(\mathcal{M}) \overset{\text{(i)}}{\leq} \frac{\tau^2}{\sqrt{m}}, \quad (5.20)$$

where (h) follows from (5.13) and (i) follows from (5.14). Finally, we use these bounds on U and V to bound the quantity $2\exp\big(-c_2 \min\{\frac{t^2}{V^2}, \frac{t}{U}\}\big)$as

$$\begin{aligned}
2\exp\Big(-c_2 \min\Big\{\frac{t^2}{V^2}, \frac{t}{U}\Big\}\Big) &\leq 2\exp\Big(-c_2 \min\Big\{m\Big(\frac{t}{\tau^2}\Big)^2, \frac{tm}{\tau^2}\Big\}\Big) \\
&\overset{\text{(j)}}{\leq} 2\exp\Big(-c_2 \min\Big\{\Big(\frac{\delta}{2\tau^2}\Big)^2 \frac{K_2\,\log(\varepsilon^{-1})}{\delta^2}, \frac{K_2\,\log(\varepsilon^{-1})}{2\tau^2\,\delta}\Big\}\Big) \\
&= 2\exp\Big(-\frac{c_2\,K_2\,\log(\varepsilon^{-1})}{2\delta} \min\Big\{\frac{\delta}{2\tau^4}, \frac{1}{\tau^2}\Big\}\Big) \overset{\text{(k)}}{\leq} \varepsilon, \quad (5.21)
\end{aligned}$$

where (j) follows from setting $t = \frac{\delta}{2}$ and using the bound on m, while (k) holds true for

$$K_2 \geq \max\Big\{(2\tau^2)^2, 2\delta\tau^2\Big\}\Big(\frac{\log(1/2)}{c_2\log(\varepsilon)} + \frac{1}{c_2}\Big).$$

Using (5.18), (5.21), and $t = \frac{\delta}{2}$ with Theorem 5.2, the proof of this theorem follows. $\square$

5.9 Auxiliary Lemmas

Lemma 5.4 ([84]). *For any fixed notion of norm $\|\,.\,\|$, define a unit-norm ball $\mathcal{B}_1 := \{x \in \mathbb{R}^n : \|x\| \leq 1\}$ with distance measure $\|\,.\,\|$. Then the covering number of $\mathcal{B}_1$ (with respect to the norm $\|\,.\,\|$) satisfies the bound*

$$\Psi(\mathcal{B}_1, \|\,.\,\|, \epsilon) \leq \left(\frac{3}{\epsilon}\right)^n, \epsilon \in (0,1).$$

Lemma 5.5 ([90]). *Define metric spaces $(\mathcal{D}_1, h_1)$, $(\mathcal{D}_2, h_2)$, ..., $(\mathcal{D}_p, h_p)$. Further, define the Cartesian product $\mathcal{D}_0 := \mathcal{D}_1 \times_1 \mathcal{D}_2 \times_2 \cdots \times_p \mathcal{D}_p$ with respect to the norm $h_0(D_0^1, D_0^2) = \max_{j \in [[p]]} \left\{h_j(D_j^1, D_j^2)\right\}$, where $D_0^1, D_0^2 \in \mathcal{D}_0$ such that $D_0^1 = D_1^1 \times_1 D_2^1 \times_2 \cdots \times_p D_p^1$, $D_0^2 = D_1^2 \times_1 D_2^2 \times_2 \cdots \times_p D_p^2$, and $D_j^1, D_j^2 \in \mathcal{D}_j$ for any $j \in [[p]]$. Then the covering number of $\mathcal{D}_0$ (with respect to the norm h_0) satisfies the bound*

$$\Psi(\mathcal{D}_0, h_0, \epsilon) \leq \prod_{j=1}^{d} \Psi(\mathcal{D}_j, h_j, \epsilon).$$

Lemma 5.6 ([91]). *Define sets $\mathcal{D}_1$ and $\mathcal{D}_2$ with distance measures h_1 and h_2, respectively. Further, define map $\Phi : \mathcal{K} \to \mathcal{D}_2$ such that $\mathcal{K} \subset \mathcal{D}_1$. Then, for some $L > 0$, if Φ satisfies*

$$h_2(\Phi(K_1), \Phi(K_2)) \leq L\, h_1(K_1, K_2) \;\; \text{for } K_1, K_2 \in \mathcal{K},$$

i.e. Φ is a Lipschitz map with constant L, then, for any $\epsilon > 0$, we have

$$\Psi(\Phi(\mathcal{K}), h_2, L\,\epsilon) \leq \Psi(\mathcal{K}, h_1, \epsilon).$$

Bibliography

[1] T. R. Golub et al., "Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring," *Science*, vol. 286, no. 5439, pp. 531–537, 1999.

[2] X. Huang and W. Pan, "Linear regression and two-class classification with gene expression data," *Bioinformatics*, vol. 19, no. 16, pp. 2072–2078, 2003.

[3] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Foundations and Trends in Information Retrieval*, vol. 2, no. 1–2, pp. 1–135, 2008.

[4] B. Liu, "Sentiment analysis and opinion mining," *Synthesis Lectures on Human Language Technologies*, vol. 5, no. 1, pp. 1–167, 2012.

[5] D. Landgrebe, "Hyperspectral image data analysis," *IEEE Signal Processing Mag.*, vol. 19, no. 1, pp. 17–28, 2002.

[6] A. Plaza et al., "Recent advances in techniques for hyperspectral image processing," *Remote Sens. Environ.*, vol. 113, pp. S110–S122, 2009.

[7] J. Bioucas-Dias et al., "Hyperspectral unmixing overview: Geometrical, statistical, and sparse regression-based approaches," *IEEE J. Select. Topics Appl. Earth Observ. Remote Sensing*, vol. 5, no. 2, pp. 354–379, 2012.

[8] M. A. Lindquist *et al.*, "The statistical analysis of fMRI data," *Statistical science*, vol. 23, no. 4, pp. 439–464, 2008.

[9] C. Hinrichs, V. Singh, L. Mukherjee, G. Xu, M. K. Chung, S. C. Johnson, A. D. N. Initiative *et al.*, "Spatially augmented LPboosting for AD classification with evaluations on the ADNI dataset," *Neuroimage*, vol. 48, no. 1, pp. 138–149, 2009.

[10] S. Gandy, B. Recht, and I. Yamada, "Tensor completion and low-n-rank tensor recovery via convex optimization," *Inverse Problems*, vol. 27, no. 2, p. 025010, 2011.

[11] E. J. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. Inform. Theory*, vol. 52, no. 2, pp. 489–509, 2006.

[12] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning*. Springer Texts in Statistics, 2013.

[13] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Roy. Statist. Soc. B.*, vol. 58, no. 1, pp. 267–288, 1996.

[14] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *J. Roy. Statist. Soc.*, vol. 67, no. 2, pp. 301–320, 2005.

[15] N. Li and B. Li, "Tensor completion for on-board compression of hyperspectral images," in *Proc. Intl. Conf. Image Process (ICIP)*. IEEE, 2010, pp. 517–520.

[16] T. G. Kolda and B. W. Bader, "Tensor decompositions and applications," *SIAM Review*, vol. 51, no. 3, pp. 455–500, 2009.

[17] B. Romera-Paredes, H. Aung, N. Bianchi-Berthouze, and M. Pontil, "Multilinear multitask learning," in *Proc. Intl. Conf. Mach. Learning (ICML)*, 2013, pp. 1444–1452.

[18] D. Nion and N. D. Sidiropoulos, "Tensor algebra and multidimensional harmonic retrieval in signal processing for MIMO radar," *IEEE Trans. Signal Process.*, vol. 58, no. 11, pp. 5693–5705, 2010.

[19] H. Wang and M. Thoss, "Numerically exact quantum dynamics for indistinguishable particles: The multilayer multiconfiguration time-dependent hartree theory in second quantization representation," *The J. Chemical Physics*, vol. 131, no. 2, p. 024114, 2009.

[20] N. D. Sidiropoulos, L. De Lathauwer, X. Fu, K. Huang, E. E. Papalexakis, and C. Faloutsos, "Tensor decomposition for signal processing and machine learning," *IEEE Trans. Signal Process.*, vol. 65, no. 13, pp. 3551–3582, 2017.

[21] R. Tomioka, T. Suzuki, K. Hayashi, and H. Kashima, "Statistical performance of convex tensor decomposition," in *Proc. Advances in Neural Inform. Process. Systems (NIPS)*, 2011, pp. 972–980.

[22] H. Rauhut, R. Schneider, and Ž. Stojanac, "Low rank tensor recovery via iterative hard thresholding," *Linear Algebra and its Applications*, vol. 523, pp. 220–262, 2017.

[23] C. Mu, B. Huang, J. Wright, and D. Goldfarb, "Square deal: Lower bounds and improved relaxations for tensor recovery," in *Proc. Intl. Conf. Mach. Learning (ICML)*, 2014, pp. 73–81.

[24] P. Bellec, C. Chu, F. Chouinard-Decorte, Y. Benhajali, D. S. Margulies, and R. C. Craddock, "The neuro bureau ADHD-200 preprocessed repository," *Neuroimage*, vol. 144, pp. 275–286, 2017.

[25] D. L. Donoho, "For most large underdetermined systems of linear equations the minimal ℓ_1-norm solution is also the sparsest solution," *Commun. Pure Appl. Math.*, vol. 59, no. 6, pp. 797–829, 2006.

[26] C. R. Genovese, J. Jin, L. Wasserman, and Z. Yao, "A comparison of the lasso and marginal regression," *J. Machine Learning Res.*, vol. 13, pp. 2107–2143, 2012.

[27] B. Hao, A. Zhang, and G. Cheng, "Sparse and low-rank tensor estimation via cubic sketchings," *arXiv preprint arXiv:1801.09326*, 2018.

[28] G. Raskutti, M. Yuan, and H. Chen, "Convex regularization for high-dimensional multi-response tensor regression," *The Annals of Statistics*, vol. 47, no. 3, pp. 1554–1584, 2019.

[29] J. Fan and R. Li, "Variable selection via nonconcave penalized likelihood and its oracle properties," *J. Am. Statist. Ass.*, vol. 96, no. 456, pp. 1348–1360, 2001.

[30] W. J. Fu, "Penalized regressions: The bridge versus the lasso," *J. Computnl Graph. Statist.*, vol. 7, no. 3, pp. 397–416, 1998.

[31] J. Huang, J. L. Horowitz, and S. Ma, "Asymptotic properties of bridge estimators in sparse high-dimensional regression models," *Ann. Stat.*, vol. 36, no. 2, pp. 587–613, 2008.

[32] H. Zou, "The adaptive lasso and its oracle properties," *J. Am. Statist. Ass.*, vol. 101, no. 476, pp. 1418–1429, 2006.

[33] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *J. Royal Statistical Soc. B*, vol. 68, no. 1, pp. 49–67, 2006.

[34] E. Candes and T. Tao, "The Dantzig selector: Statistical estimation when p is much larger than n," *Ann. Stat.*, vol. 35, no. 6, pp. 2313–2351, 2007.

[35] D. L. Donoho, "High-dimensional data analysis: The curses and blessings of dimensionality," *AMS Math Challenges Lecture*, vol. 1, p. 32, 2000.

[36] J. Fan and J. Lv, "Sure independence screening for ultrahigh dimensional feature space," *J. Roy. Statist. Soc. B.*, vol. 70, no. 5, pp. 849–911, 2008.

[37] J. Fan, R. Samworth, and Y. Wu, "Ultrahigh dimensional feature selection: Beyond the linear model," *J. Machine Learning Res.*, vol. 10, pp. 2013–2038, 2009.

[38] J. Fan and R. Song, "Sure independence screening in generalized linear models with NP-dimensionality," *Ann. Stat.*, vol. 38, no. 6, pp. 3567–3604, 2010.

[39] J. Fan, Y. Feng, and R. Song, "Nonparametric independence screening in sparse ultra-high dimensional additive models," *J. Am. Statist. Ass.*, vol. 106, no. 494, pp. 544–557, 2011.

[40] J. Fan, Y. Ma, and W. Dai, "Nonparametric independence screening in sparse ultra-high dimensional varying coefficient models," *J. Am. Statist. Ass.*, vol. 109, no. 507, pp. 1270–1284, 2014.

[41] J. Fan, Y. Feng, and Y. Wu, "High-dimensional variable selection for Cox's proportional hazards model," in *Borrowing Strength: Theory Powering Applications–A Festschrift for Lawrence D. Brown.* Institute of Mathematical Statistics, 2010, pp. 70–86.

[42] S. D. Zhao and Y. Li, "Principled sure independence screening for Cox models with ultra-high-dimensional covariates," *J. Multivariate Anal.*, vol. 105, no. 1, pp. 397–411, 2012.

[43] L. E. Ghaoui, V. Viallon, and T. Rabbani, "Safe feature elimination for the lasso and sparse supervised learning problems," *arXiv preprint arXiv:1009.4219*, 2010.

[44] R. Tibshirani et al., "Strong rules for discarding predictors in lasso-type problems," *J. Roy. Statist. Soc. B.*, vol. 74, no. 2, pp. 245–266, 2012.

[45] Z. J. Xiang and P. J. Ramadge, "Fast lasso screening tests based on correlations," in *Proc. IEEE Int. Conf. on Acoustics Speech and Sig. Proc. (ICASSP)*, 2012, pp. 2137–2140.

[46] L. Dai and K. Pelckmans, "An ellipsoid based, two-stage screening test for BPDN," in *Proc. 20th Eur. Sig. Proc. Conf.*, 2012, pp. 654–658.

[47] J. Wang, P. Wonka, and J. Ye, "Lasso screening rules via dual polytope projection," *J. Mach. Learn. Res.*, vol. 16, pp. 1063–1101, 2015.

[48] P. Hall and H. Miller, "Using generalized correlation to effect variable selection in very high dimensional problems," *J. Computnl Graph. Statist.*, vol. 18, no. 3, pp. 533–550, 2009.

[49] G. Li, H. Peng, J. Zhang, and L. Zhu, "Robust rank correlation based screening," *Ann. Stat.*, vol. 40, no. 3, pp. 1846–1877, 2012.

[50] W. U. Bajwa, R. Calderbank, and S. Jafarpour, "Why Gabor frames? Two fundamental measures of coherence and their role in model selection," *J. Commun. Netw.*, vol. 12, no. 4, pp. 289–307, 2010.

[51] G. Davis, S. Mallat, and M. Avellaneda, "Adaptive greedy approximations," *J. Construct. Approx.*, vol. 13, no. 1, pp. 57–98, 1997.

[52] M. J. Wainwright, "Sharp thresholds for high-dimensional and noisy sparsity recovery using-constrained quadratic programming (lasso)," *IEEE Trans. Inf. Theory*, vol. 55, no. 5, pp. 2183–2202, 2009.

[53] P. Zhao and B. Yu, "On model selection consistency of lasso," *J. Machine Learning Res.*, vol. 7, pp. 2541–2563, 2006.

[54] E. J. Candes, "The restricted isometry property and its implications for compressed sensing," *Comptes Rendus Mathematique*, vol. 346, no. 9-10, pp. 589–592, 2008.

[55] G. Raskutti, M. J. Wainwright, and B. Yu, "Restricted eigenvalue properties for correlated Gaussian designs," *J. Machine Learning Res.*, vol. 11, pp. 2241–2259, 2010.

[56] W. U. Bajwa, A. Calderbank, and D. G. Mixon, "Two are better than one: Fundamental parameters of frame coherence," *Appl. Comput. Harmon. Anal.*, vol. 33, pp. 58–78, 2012.

[57] L. Welch, "Lower bounds on the maximum cross correlation of signals," *IEEE Trans. Inform. Theory*, vol. 20, no. 3, pp. 397–399, 1974.

[58] A. L. Maas et al., "Learning word vectors for sentiment analysis," in *Proc. 49th Ann. Meeting of the Assoc. for Computational Linguistics: Human Language Technologies*, June 2011, pp. 142–150.

[59] J. Fan and Y. Fan, "High dimensional classification using features annealed independence rules," *Ann. Stat.*, vol. 36, no. 6, p. 2605, 2008.

[60] J. Fan, F. Han, and H. Liu, "Challenges of big data analysis," *Nat. Sci. Rev.*, vol. 1, no. 2, pp. 293–314, 2014.

[61] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval.* Cambridge University Press, 2008.

[62] J. Wainwright, "High-dimensional statistics: A non-asymptotic viewpoint," *In preparation. University of California, Berkeley*, 2015. [Online]. Available: https://www.stat.berkeley.edu/~mjwain/stat210b/Chap2_TailBounds_Jan22_2015.pdf

[63] N. Lazar, *The statistical analysis of functional MRI data*. Springer Science & Business Media, 2008.

[64] S. Ryali, K. Supekar, D. A. Abrams, and V. Menon, "Sparse logistic regression for whole-brain classification of fMRI data," *NeuroImage*, vol. 51, no. 2, pp. 752–764, 2010.

[65] B. S. Caffo, C. M. Crainiceanu, G. Verduzco, S. Joel, S. H. Mostofsky, S. S. Bassett, and J. J. Pekar, "Two-stage decompositions for the analysis of functional connectivity for fMRI with application to Alzheimer's disease risk," *NeuroImage*, vol. 51, no. 3, pp. 1140–1149, 2010.

[66] T. Ahmed and W. U. Bajwa, "Exsis: Extended sure independence screening for ultrahigh-dimensional linear models," *Signal Processing*, vol. 159, pp. 33–48, 2019.

[67] Z. J. Xiang, Y. Wang, and P. J. Ramadge, "Screening tests for lasso problems," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 5, pp. 1008–1027, 2017.

[68] J. Liu, P. Musialski, P. Wonka, and J. Ye, "Tensor completion for estimating missing values in visual data," in *Proc. Intl .Conf. Computer Vision (ICCV)*.

[69] T. Blumensath and M. E. Davies, "Iterative hard thresholding for compressed sensing," *Applied and Computational Harmonic Analysis*, vol. 27, no. 3, pp. 265–274, 2009.

[70] R. Garg and R. Khandekar, "Gradient descent with sparsification: an iterative algorithm for sparse recovery with restricted isometry property," in *Proc. Intl. Conf. Mach. Learning (ICML)*. ACM, 2009, pp. 337–344.

[71] P. Jain, R. Meka, and I. S. Dhillon, "Guaranteed rank minimization via singular value projection," in *Advances in Neural Inform. Process. Systems (NIPS)*, 2010, pp. 937–945.

[72] R. Yu and Y. Liu, "Learning from multiway data: Simple and efficient tensor regression," in *Intl. Conf. Machine Learning (ICML)*, 2016, pp. 373–381.

[73] H. Chen, G. Raskutti, and M. Yuan, "Non-convex projected gradient descent for generalized low-rank tensor regression," *arXiv preprint arXiv:1611.10349*, 2016.

[74] I. Rish and G. Grabarnik, *Sparse Modeling: Theory, Algorithms, and Applications.* CRC press, 2014.

[75] C. J. Hillar and L.-H. Lim, "Most tensor problems are NP-hard," *Journal of the ACM (JACM)*, vol. 60, no. 6, p. 45, 2013.

[76] G. Allen, "Sparse higher-order principal components analysis," in *Proc. Artificial Intelligence and Stat. (AISTATS)*, 2012, pp. 27–36.

[77] L. Grasedyck, D. Kressner, and C. Tobler, "A literature survey of low-rank tensor approximation techniques," *GAMM-Mitteilungen*, vol. 36, no. 1, pp. 53–78, 2013.

[78] W. W. Sun, J. Lu, H. Liu, and G. Cheng, "Provable sparse tensor decomposition," *J. Royal Statistical Society: Series B (Statistical Methodology)*, vol. 79, no. 3, pp. 899–916, 2017.

[79] M. Hein and T. Bühler, "An inverse power method for nonlinear eigenproblems with applications in 1-spectral clustering and sparse PCA," in *Proc. Advances in Neural Information Processing Systems*, 2010, pp. 847–855.

[80] H. Zhou, L. Li, and H. Zhu, "Tensor regression with applications in neuroimaging data analysis," *J. American Statistical Association*, vol. 108, no. 502, pp. 540–552, 2013.

[81] N. Vervliet, O. Debals, L. Sorber, M. Van Barel, and L. De Lathauwer, "Tensorlab 3.0, Mar. 2016," *URL http://www.tensorlab.net. Available online.*

[82] B. W. Bader et al., "Matlab tensor toolbox version 3.0-dev," Available online, Oct. 2017. [Online]. Available: https://www.tensortoolbox.org

[83] K. D. Ba, P. Indyk, E. Price, and D. P. Woodruff, "Lower bounds for sparse recovery," in *Proceedings of the twenty-first annual ACM-SIAM symposium on Discrete Algorithms.* SIAM, 2010, pp. 1190–1197.

[84] E. J. Candes and Y. Plan, "Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements," *IEEE Transactions on Information Theory*, vol. 57, no. 4, pp. 2342–2359, 2011.

[85] F. Krahmer, S. Mendelson, and H. Rauhut, "Suprema of chaos processes and the restricted isometry property," *Communications on Pure and Applied Mathematics*, vol. 67, no. 11, pp. 1877–1904, 2014.

[86] M. Talagrand, *Upper and Lower Bounds for Stochastic Processes: Modern Methods and Classical Problems*. Springer Science & Business Media, 2014, vol. 60.

[87] M. P. Milham, D. Fair, M. Mennes, S. H. Mostofsky *et al.*, "The ADHD-200 consortium: a model to advance the translational potential of neuroimaging in clinical neuroscience," *Frontiers in Systems Neuroscience*, vol. 6, p. 62, 2012.

[88] Q.-H. Zou, C.-Z. Zhu, Y. Yang, X.-N. Zuo, X.-Y. Long, Q.-J. Cao, Y.-F. Wang, and Y.-F. Zang, "An improved approach to detection of amplitude of low-frequency fluctuation (ALFF) for resting-state fMRI: fractional ALFF," *J. Neuroscience Methods*, vol. 172, no. 1, pp. 137–141, 2008.

[89] H. Yang, Q.-Z. Wu, L.-T. Guo, Q.-Q. Li, X.-Y. Long, X.-Q. Huang, R. C. Chan, and Q.-Y. Gong, "Abnormal spontaneous brain activity in medication-naive ADHD children: A resting state fMRI study," *Neuroscience Letters*, vol. 502, no. 2, pp. 89–93, 2011.

[90] R. Gribonval, R. Jenatton, F. Bach, M. Kleinsteuber, and M. Seibert, "Sample complexity of dictionary learning and other matrix factorizations," *IEEE Transactions on Information Theory*, vol. 61, no. 6, pp. 3469–3486, 2015.

[91] S. Szarek, "Metric entropy of homogeneous spaces," *Banach Center Publications*, vol. 43, no. 1, pp. 395–410, 1998.