Distributed Feature Selection in Large $n$ and Large $p$ Regression Problems

by

Xiangyu Wang

Department of Statistical Science
Duke University

Date: ________________
Approved: ________________

David B. Dunson, Supervisor

Surya T. Tokdar

Katherine A. Heller

Galen Reeves

Dissertation submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in the Department of Statistical Science
in the Graduate School of Duke University
2016
Abstract

Distributed Feature Selection in Large $n$ and Large $p$ Regression Problems

by

Xiangyu Wang

Department of Statistical Science
Duke University

Date: ____________________

Approved:

____________________________
David B. Dunson, Supervisor

____________________________
Surya T. Tokdar

____________________________
Katherine A. Heller

____________________________
Galen Reeves

An abstract of a dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Statistical Science in the Graduate School of Duke University 2016
Abstract

Fitting statistical models is computationally challenging when the sample size or the dimension of the dataset is huge. An attractive approach for down-scaling the problem size is to first partition the dataset into subsets and then fit using distributed algorithms. The dataset can be partitioned either horizontally (in the sample space) or vertically (in the feature space), and the challenge arise in defining an algorithm with low communication, theoretical guarantees and excellent practical performance in general settings. For sample space partitioning, I propose a MEdian Selection Subset AGgregation Estimator (message) algorithm in answering to the challenges. The algorithm applies feature selection in parallel for each subset using regularized regression or Bayesian variable selection method, calculates the ‘median’ feature inclusion index, estimates coefficients for the selected features in parallel for each subset, and then averages these estimates. The algorithm is simple, involves very minimal communication, scales efficiently in sample size, and has theoretical guarantees. I provide extensive experiments to show excellent performance in feature selection, estimation, prediction, and computation time relative to usual competitors.

While sample space partitioning is useful in handling datasets with large sample size, feature space partitioning is more effective when the data dimension is high. Existing methods for partitioning features, however, are either vulnerable to high correlations or inefficient in reducing the model dimension. In the thesis, I propose a new embarrassingly parallel framework named DECO for distributed variable
selection and parameter estimation. In DECO, variables are first partitioned and allocated to m distributed workers. The decorrelated subset data within each worker are then fitted via any algorithm designed for high-dimensional problems. We show that by incorporating the decorrelation step, DECO can achieve consistent variable selection and parameter estimation on each subset with (almost) no assumptions. In addition, the convergence rate is nearly minimax optimal for both sparse and weakly sparse models and does NOT depend on the partition number m. Extensive numerical experiments are provided to illustrate the performance of the new framework.

For datasets with both large sample sizes and high dimensionality, I propose a new “divided-and-conquer” framework DEME (DECO-message) by leveraging both the DECO and the message algorithm. The new framework first partitions the dataset in the sample space into row cubes using message and then partition the feature space of the cubes using DECO. This procedure is equivalent to partitioning the original data matrix into multiple small blocks, each with a feasible size that can be stored and fitted in a computer in parallel. The results are then synthesized via the DECO and message algorithm in a reverse order to produce the final output. The whole framework is extremely scalable.
To my family.
Contents

Abstract iv
List of Tables x
List of Figures xi
List of Abbreviations and Symbols xii
Acknowledgements xiii
1 Introduction 1
2 Feature selection in a nutshell 7
  2.1 Overview . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 7
  2.2 Regularized regression . . . . . . . . . . . . . . . . . . . . . . . . . 9
    2.2.1 The loss function and optimization . . . . . . . . . . . . . . . 9
    2.2.2 The non-asymptotic theory . . . . . . . . . . . . . . . . . . . 10
  2.3 Bayesian variable selection . . . . . . . . . . . . . . . . . . . . . . . 12
    2.3.1 Prior specification and sampling . . . . . . . . . . . . . . . . 13
    2.3.2 The non-asymptotic theory on posterior concentration . . . . 15
  2.4 Preconditioning and the elliptical family . . . . . . . . . . . . . . . 16
    2.4.1 Preconditioning for high dimensional linear regression . . . . 16
    2.4.2 The elliptical family . . . . . . . . . . . . . . . . . . . . . . 18
3 Sample space partition using message 21
  3.1 Introduction . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 21
3.2 The message algorithm ................................................. 23
3.3 Consistency and robustness ........................................... 25
  3.3.1 Convergence rate for regularized regression ................... 26
  3.3.2 Convergence rate for Bayesian variable selection ............... 30
3.4 Experiments .............................................................. 35
  3.4.1 Synthetic data sets .................................................. 35
  3.4.2 Individual household electric power consumption ............... 40
  3.4.3 HIGGS classification ............................................... 41

4 Feature space partition using DECO ................................. 42
  4.1 Introduction ............................................................. 42
  4.2 The DECO framework .................................................. 44
  4.3 Theoretical properties ............................................... 48
    4.3.1 The convergence rate for regularized regression ............... 48
    4.3.2 The convergence rate for Bayesian variable selection ......... 52
  4.4 The relationship with variable screening .......................... 57
  4.5 Experiments ............................................................. 59
    4.5.1 Monitor DECO on one subset .................................... 61
    4.5.2 Impact of the subset number $m$ ................................ 62
    4.5.3 Comprehensive comparison ..................................... 63
    4.5.4 Real data .......................................................... 64

5 DEME: A block-wise partitioning scheme ......................... 68
  5.1 Introduction ............................................................. 68
  5.2 The DEME framework .................................................. 69
  5.3 The consistency theory ............................................... 72

6 Conclusion and future extensions ................................. 74
A Theory on the elliptical family

A.1 Proof for Theorem 2.4 ............................................. 76
A.2 Proofs for Theorem 2.5 and 2.6 ................................. 86
A.3 Proof of Lemma 3 .................................................. 92

B Theory on the regularized regression ......................... 94

C Theory on the Bayesian variable selection .................. 101

C.1 Deterministic conditions for Bayesian variable selection . 101
C.2 Relating to Condition 2 ........................................... 106

Bibliography ............................................................ 111

Biography ............................................................... 115
## List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1</td>
<td>Results for five models with $(n, p) = (500, 10000)$</td>
<td>64</td>
</tr>
<tr>
<td>4.2</td>
<td>The results of all methods for student performance data with $(n, p, m) = (395, 741, 5)$</td>
<td>65</td>
</tr>
<tr>
<td>4.3</td>
<td>The results of all methods for mammalian eye diseases with $(n, p, m) = (120, 5000, 100)$</td>
<td>66</td>
</tr>
<tr>
<td>4.4</td>
<td>The results of all methods for electricity load diagram data with $(n, p, m) = (370, 1251980, 1000)$</td>
<td>67</td>
</tr>
</tbody>
</table>
List of Figures

1.1 Illustration of the sample space partition . . . . . . . . . . . . . . . . 3
1.2 Illustration of the feature space partition . . . . . . . . . . . . . . . . 5
3.1 Results for case 1 with $\rho = 0$. Top: $p = 1000$; Bottom: $p = 10,000$ . 36
3.2 Results for case 1 with $\rho = 0.5$. Top: $p = 1000$; Bottom: $p = 10,000$ . 37
3.3 Results for case 2 with $\rho = 0$. Top: $p = 1000$; Bottom: $p = 10,000$ . 37
3.4 Results for case 2 with $\rho = 0.5$. Top: $p = 1000$; Bottom: $p = 10,000$ . 38
3.5 Results for case 3 with $\rho = 0$. Top: $p = 1000$; Bottom: $p = 10,000$ . 38
3.6 Results for case 3 with $\rho = 0.5$. Top: $p = 1000$; Bottom: $p = 10,000$ . 39
3.7 Results for power consumption data. . . . . . . . . . . . . . . . . . . . 40
3.8 Results for HIGGS classification. . . . . . . . . . . . . . . . . . . . . 41
4.1 Performance of DECO on one subset with $p = 10,000$ and different $n$’s. 61
4.2 Performance of DECO on one subset with $n = 500$ and different $p$’s. . 62
4.3 Performance of DECO with different number of subsets. . . . . . . . 63
5.1 The partitioning scheme for DEME . . . . . . . . . . . . . . . . . . . . 69
5.2 Implement DEME on the GPU. . . . . . . . . . . . . . . . . . . . . . . 71
5.3 Implement DEME in mapreduce framework. . . . . . . . . . . . . . . 71
List of Abbreviations and Symbols

Symbols

\begin{itemize}
\item \(\mathbb{E}\) The expectation operation
\item \(\mathbb{P}\) The probability symbol
\item \(\mathcal{R}\) The set of real numbers
\item \(e_i\) The unit vector with \(i^{th}\) coordinate being 1.
\item \(\chi^2(n)\) The chi-square random variable with degree \(n\)
\item \([p]\) The set of \(\{1, 2, \cdots, p\}\)
\item \(N(0, \Sigma)\) The multivariate Gaussian random variable with mean zero and covariance \(\Sigma\)
\item \(I_p\) The \(p \times p\) identity matrix
\end{itemize}

Abbreviations

\begin{itemize}
\item \textit{lasso} Least absolute shrinkage and selection operator (Tibshirani, 1996)
\item \textit{SCAD} Smoothly clipped absolute deviation (Fan and Li, 2001)
\item \textit{MCP} Minimax concavity penalty (Zhang, 2010)
\item \textit{HOLP} High dimensional ordinary least squares projection (Wang and Leng, 2015)
\item \textit{MACG} Matrix angular central distribution (Chikuse, 2003)
\end{itemize}
Acknowledgements

I felt myself being so blessed for the past four years. I fortunately met a lot of incredible people, making my Ph.D. life an invaluable memory. I would like to express my deep gratefulness to all these people. First of all, I want to thank my advisor, Dr. David Dunson, for being a great mentor and friend to me. David is a great statistician of this generation, not only because of his wisdom and creativeness, but also his great vision. His way of approaching problems and new ideas shaped my whole research style, from which I benefit tremendously during my Ph.D. study and, I believe, will continue to for my future life.

I thank Dr. Chenlei Leng for leading me into the world of statistics. I met Chenlei when I was undergrad and we collaborated on multiple papers since then. Chenlei can easily see things beyond my perspective and guide me through the difficulties. Although not officially on record, I considered Chenlei as my co-advisor and appreciate all the help during my Ph.D. life.

I thank Dr. Katherine Heller for all the supports and help. Katherine was my advisor for the master’s degree in computer science and guided me on my very first project at Duke. We started collaborating since then.

I would also like to thank my other thesis committee members, Dr. Surya Tokdar, Dr. Galen Reeves and Dr. Rong Ge for all your constructive suggestions and discussions on my research work. I want to thank Dr. Mike West for all your help including the long email that directed me to the Duke Stat department. This is the
most important and best ever decision I’ve made in my life. I also thank Dr. Tamraparni Dasu and Dr. Horace Ip for the guidance on my summer project at AT&T Labs and Christopher Suter and Monica Lenart for my summer project at Google.

Many special thanks go to my fellows and colleagues. I want to thank Fangjian Guo for being a great co-author and amazing friend during my stay at Duke. I very much enjoyed our discussions and brainstormings, and all the frustrations when figuring out we were sometimes re-inventing the ”old wheels”. I want to thank Yizhe Zhang, Kai Fan and Dr. Peichao Peng for all the research collaborations, from which I benefit a lot. I would like to thank all the extremely talented people in the Dunson’ Group and Heller’s group for numerous conversations and discussions that inspired my research.

I am deeply grateful for all the fun time I had with my friends during my Ph.D. study. I thank Dr. Fangpo Wang and Dr. Yun Yang for all the interesting discussions and movie time. I thank Dr. Jianling Zhong and Dr. Jiangwei Pan for being great roommates and swim friends and Dr. Jingchen Hu and Shuting (Lise) Lv for all the fun time we had on Japanese animations and mangas. I appreciate all the cultural discussions and game play with Dr. Kaoru Irie and Michael Lindon and I thank all the fellows in 017 Old Chem for the good time. I would like to thank everyone in the department for their invaluable help and all my friends for making my Ph.D. life cheerful.

My research was supported by the National Institutes of Health (NIH R01-ES017436). I would like to thank the Section on Bayesian Statistical Science for funding my travel to JSM 2014. I would also like to thank the Neural Information Processing System foundation and the International Machine Learning Society for funding my travel to NIPS 2014, NIPS 2015 and ICML 2016.

Finally, I want to express my deepest thankfulness to my parents. I thank everything they have done for me and filling my life with joy.
Introduction

The explosion in both size and dimensionality of data has brought new challenges to the design of modern statistical algorithms. The challenges typically lie in two aspects. First, the computational time increases dramatically due to the enormous amount of data processed at each iteration of an algorithm. Second, the storage and the memory size face huge burdens for storing and accessing the large datasets. To ameliorate these issues, many methods such as the sub-sampling and the stochastic gradient descent have been exploited, among which the parallel computing serves as the most promising approach. The typical procedure for parallelization partitions the full data into multiple subsets, stores subsets on different machines, and then processes subsets simultaneously. Processing on subsets in parallel can lead to two types of computational gains. The first reduces time for calculations within each iteration of optimization or sampling algorithms via faster operations; for example, in conducting linear algebra involved in calculating likelihoods or gradients. Although such approaches can lead to substantial reductions in computational bottlenecks for big data, the amount of gain is limited by the need to communicate across computers at each iteration. It is well known that communication costs are a major factor
driving the efficiency of distributed algorithms, so that it is of critical importance to limit communication. This motivates the second type of approach, the embarrassingly parallelization, which conducts computations completely independently on the different subsets, and then combines the results to obtain the final output. This limits communication to the final combining step, and may lead to simpler and faster algorithms. However, a major issue for this approach is how to design algorithms that are close to communication free, but can preserve or even improve the statistical accuracy relative to (much slower) algorithms applied to the entire data set. I focus on addressing this challenge with an emphasize on feature selection problems in this thesis.

The successful design of an embarrassingly parallel framework for feature selection requires careful devising on two components: the data partition scheme and the subset aggregation rule. A good data partition scheme ensures a rapid computational time and efficient memory usage on each subset while a good subset aggregation rule guarantees the accuracy and the consistency of the procedure. Typically, a data matrix can be partitioned either horizontally (in the sample space, see Figure 1.1) or vertically (in the feature space, see Figure 1.2). The sample space partitioning often occurs in the so-called "large n" problem where the sample size $n$ is huge. The advantage in using this scheme is that the samples are independent, leading to the independence among the subsets and subset-estimators after partitioning. This enables the possibility of using straightforward aggregation rule such as averaging or median to combine the subset-estimators while maintaining an excellent accuracy. There is a recent flurry of research in both Bayesian and frequentist methods focusing on this partition scheme (Zhang et al., 2012; Mann et al., 2009; Scott et al., 2013; Neiswanger et al., 2013; Wang and Dunson, 2013; Minsker et al., 2015, 2014). In Chapter 3, we extends the median approach to feature selection and propose a parallel feature selection algorithm (message) based on sample space partitioning and median
selection aggregation. The *message* algorithm first randomly partitions the dataset in the sample space and applies feature selection in parallel for each subset. The selected sub-models are then aggregated by calculating the “median” feature inclusion index (or equivalently, via majority voting) to obtain the final model. We show that the algorithm is consistent and robust and achieves similar and often times better feature selection result compared to algorithms applied directly to the full dataset.

![Figure 1.1: Illustration of the sample space partition](image)

While sample space partitioning can be useful for downscaling the big sample size, it is less appealing for problems with high dimension $p \gg n$. These problems are known as the "large $p$" problem and are routinely seen in scientific and technological applications. In contrast to sample space partitioning, feature space partitioning is more effective for dimension reduction, and is thus desired in solving "large $p$" problem. Intuitively, feature space partitioning is much more challenging than sample space partitioning, mainly because of the correlations between features. A partition
of the feature space would succeed only when the features across the partitioned subsets were mutually independent. Otherwise, it is highly likely that any model posed on the subsets is misspecified and the results are biased regardless of the sample size. In reality, however, mutually independent groups of features may not exist; Even if they do, finding these groups is likely more challenging than fitting a high-dimensional model. Therefore, although conceptually attractive, feature space partitioning is extremely challenging. On the other hand, feature space partitioning is straightforward if the features are independent. Motivated by this key fact, we propose, in Chapter 4, a novel embarrassingly-parallel framework named DECO by decorrelating the features before partitioning. In DECO, variables are first partitioned and allocated to \( m \) distributed workers. A decorrelation step is then conducted within each worker and the decorrelated data are fitted via any algorithm designed for high-dimensional problems. We show that by incorporating the decorrelation step, DECO can achieve consistent variable selection and parameter estimation on each subset with (almost) no assumptions. In addition, the convergence rate does NOT depend on the partition number \( m \). In view of the huge computational gain and the easy implementation, DECO is extremely attractive for solving ”large \( p \)” problems.

The proposed two algorithms are carefully designed for solving each side of the ”big data” problem, i.e., ”large \( p \)” or ”large \( n \)” , but might not work well for problems with both large sample size and high dimensionality. In solving the feature selection problem on ”large-n-and-large-p” datasets, we propose a new ”divided-and-conquer” framework DEME (DECO-message) in Chapter 5, by leveraging both the DECO and the message algorithm. The new framework first partitions the large-\( n \)-and-large-\( p \) dataset in the sample space to obtain \( l \) row cubes such that each becomes a large-\( p \)-small-\( n \) dataset. We then partition the feature space of each row cube into \( m \) subsets after decorrelation. This procedure is equivalent to partitioning the original
data matrix into $l \times m$ small blocks, each with a feasible size that can be stored and fitted in a computer. We then apply the DECO algorithm to the small blocks located in the same row cube to obtain sub-estimators. The last step is to apply the message method to aggregate the $l$ row cube estimators to output the final estimate. The whole framework is extremely scalable.

This thesis presents the theory and algorithmic details of message, DECO and the new framework DEME. The basic working model used in this thesis is the linear model, with possible extensions to the generalized linear model and other non-parametric models. The thesis is organized in five chapters. In Chapter 2, I briefly review popular feature selection methods using either Bayesian or Frequentist approaches as well as some new results on elliptical distributions. The details on message and DECO algorithm are discussed in Chapter 3 and 4, including the consistency theory, the concrete algorithms and the empirical performances. In Chapter
5, I present the theory and the details of the DEME framework by leveraging results from previous chapters. A conclusion and discussion on future extensions are provided in the last Chapter and all the technical proofs are deferred to the appendix.
2

Feature selection in a nutshell

2.1 Overview

Feature selection plays an important role in processing high dimensional data for scientific research and industrial applications. With an unprecedented array of large and complex data brought by the modern data collecting technology, a defining feature of the high dimensional data is that the number of variables $p$ far exceeds the number of observations $n$. As a result, the classical ordinary least-squares estimate (OLS) used for linear regression is no longer applicable due to a lack of sufficient degrees of freedom and a good estimation on all coefficients is impossible. This motivates the idea of feature selection, i.e., to select and monitor a small group of features that have significant impact on the response and discard the rest. Such approach has been proved to be extremely useful from both the theoretical and the practical perspectives.

Consider the linear model

$$y = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon,$$

where $x = (x_1, x_2, \cdots, x_p)$ is the $p$ features, $y$ is the response, $\beta = (\beta_1, \beta_2, \cdots, \beta_p)^T$
is the coefficient and $\varepsilon$ is the noise. Alternatively, with $n$ realization of $x$ and $y$, we can rewrite the model in the matrix form

$$Y = X\beta + \varepsilon,$$

where $Y \in \mathbb{R}^n$, $X$ is an $n \times p$ matrix and $\varepsilon$ is the i.i.d noise with $\mathbb{E}[\varepsilon_i] = 0$ and $\text{var}(\varepsilon_i) = \sigma^2$. As mentioned in the beginning of this chapter, the fundamental assumption in feature selection is that the coefficients of the features can be divided into two disjoint subsets: the strong and the weak signals, where the signal strength is measured by the absolute value of the coefficients. In particular, letting $\beta^*_s$ be the coefficient of the underlying true model, the strong signal set $\gamma^*_s$ is defined as

$$\gamma^*_s = \{k : |\beta^*_sk| \geq d_*\}$$

and the weak signal set is defined as

$$(\gamma^*_s)^c = \{k : |\beta^*_sk| \leq b_*\}.$$ (2.1)

The ultimate goal of feature selection is to correctly identify the strong signal set $\gamma^*_s$ from the data. However, devising a universally good feature selection algorithm is notoriously hard (Foster et al., 2015). Just to enumerate all possible models already takes an $O(2^p)$ computational complexity without any further consideration on the selection accuracy. In general, a consistent feature selection algorithm that runs in polynomial time often relies on either the special structure of the feature matrix or the distributional property of the data.

Recent decades have witnessed an explosion in developing feature selection algorithms for handling high dimensional data sets. These methods make use of the structural information of the feature matrix with special constraints in the model space to comply the NP-dimensionality. The methods can be roughly divided into two categories in viewing of the algorithm types. The first is the regularized regression or penalized regression. See, for example, the lasso (Tibshirani, 1996), the
elastic net (Zou and Hastie, 2005), the SCAD (Fan and Li, 2001), the adaptive lasso (Zou, 2006) and the MCP (Zhang, 2010). These methods apply a regularization on the coefficient space, penalizing vectors with large norms, and shrinking smaller coefficients toward zero. Due to the continuity and the (local) convexity of the regularized loss function, any standard optimization algorithms can be employed for solving formulations in this category. The second is the Bayesian variable selection. See, for example, Yang et al. (2015). Bayesian variable selection methods allocate large portion of prior mass onto models with smaller sizes, praising sparsity in the posterior distribution. However, the discretized loss function could raise complexity issues if optimization algorithms are used. Thus, the standard inferencing algorithm is sampling, especially, the Markov Chain Monte Carlo (MCMC) algorithm.

In the remainder of this chapter, I review the basic setup and the consistency theory of the two types of feature selection algorithms (Section 2.2 and 2.3) as well as some useful technique for improving the consistency rate that will be used throughout this thesis (Section 2.4).

2.2 Regularized regression

In his section, I introduce the basic setup for regularized regression with non-concave penalties (Fan and Lv, 2011) and provide the consistency theory for these methods under suitable conditions.

2.2.1 The loss function and optimization

Recall the linear model

\[ Y = X\beta + \varepsilon. \]

The regularized regression estimates \( \hat{\beta} \) by solving the following optimization problem

\[ \hat{\beta} = \min_{\beta} \frac{1}{n} \| Y - X\beta \|_2^2 + \lambda \rho(\beta), \]

(2.3)
where $\lambda$ is a tuning parameter and $\rho_\lambda(\cdot)$ is the penalty function. The key idea for regularized regression is to impose a penalty function on the parameter space in favoring coefficients with higher sparsity. With a careful choice of the penalty norm, the solution to (2.3) could have exactly zero’s in the coordinates, fulfilling the feature selection purpose. Popular choices of the penalty functions are lasso, SCAD and MCP defined as

$$
\rho_\lambda(t) = ||t||_1, \quad \rho'_\lambda(t) = I(t \leq \lambda) + \frac{(a\lambda - t)_+}{(a-1)\lambda}I(t > \lambda), \quad \text{and} \quad \rho'_\lambda(t) = \frac{(a\lambda - t)_+}{a\lambda}
$$

For lasso, the optimization problem (2.3) is convex and can be solved efficiently using any existing convex optimization algorithms such as coordinate-wise descent (CD) or ADMM. For SCAD and MCP, the problem is non-convex. However, the convex optimization packages can still be used for searching the local sparse solution which is shown to have good feature selection property in next section. Algorithm 1 deploys how coordinate-wise descent algorithm can be used for solving (2.3).

**Algorithm 1** The coordinate-wise descent algorithm

1: Input $(Y, X, \lambda)$. Standardize $X$ and $Y$ to $x$ and $y$ with mean zero;
2: Initialize $\beta$;
3: Compute $C = X^T X/n$ and $\delta = X^T Y/n$;
4: while not converge do
5: \begin{algorithmic}
6: \State $R_i = \sum_{j \neq i} C_{ij} \beta_j$;
7: \State $\beta_i = \min C_{ii} \beta_i^2 - (2\delta_i - R_i) \beta_i + \rho_\lambda(\beta_i)$;
8: \end{algorithmic}
9: return $\beta$;

2.2.2 The non-asymptotic theory

We consider the theoretical property of $\hat{\beta}$ obtained by solving (2.3). To quantify the behavior of (2.3), one need conditions on the penalty functions and the feature matrix $X$. In particular, we cite the following condition from Fan and Lv (2011) for characterizing the general property of penalty functions.
**Condition 1** (Fan and Lv, 2009). $\rho_\lambda(t)$ is increasing and concave for $t \geq 0$, and has a continuous and increasing derivative $\rho'_\lambda(t)$ with $\rho'_\lambda(0+) = 1$. In addition, the maximum concavity $\kappa_0(\rho_\lambda)$ defined as

$$
\kappa_0(\rho_\lambda) = \sup_{0 < t_1 < t_2} \frac{-\rho'_\lambda(t_2) - \rho'_\lambda(t_1)}{t_2 - t_1}
$$

is upper bounded by some constant $\kappa_0$.

A direct calculation shows that $\kappa_0 = 0, 1/(a - 1)$ and $1/a$ for lasso, SCAD and MCP, i.e., all three penalty functions satisfy Condition 1. In addition, the following deterministic condition on $X$ is also needed.

**Condition 2.** Define $w = Y - X\beta_*$, we have

1. $\max_{i \in [p]} M_1 \leq \frac{1}{n} \|x_i\|^2_2 \leq M_2$ for some $0 < M_1 < M_2$.

2. $\max_{i \neq j \in [p]} \frac{1}{n} |x_i^T x_j| \leq \tau_1 \sigma \sqrt{\frac{\log p}{n}}$ for some $\tau_1$.

3. $\max_{i \in [p]} \frac{1}{n} |x_i^T w| \leq \tau_2 \sigma \sqrt{\frac{\log p}{n}}$ for some $\tau_2$.

Condition 2 is indeed flexible given the unspecified constants in the expression. For example, (2) is automatically satisfied if we take $\tau_1 = \sqrt{\frac{n}{\sigma^2 \log p}}$. Alternatively, one might take $\tau_1 = \frac{1}{s_*} \sqrt{\frac{n}{\sigma^2 \log p}}$, then (2) reduces to the condition in Lounici (2008).

With these two conditions, we are able to prove the following result

**Theorem 2.1.** Assume Condition 1 and 2 hold. We also assume $d_* \geq A \sigma \sqrt{\frac{\log p}{n}}$, $b_* \leq B \sigma \sqrt{\frac{\log p}{n}}$, $s_* \leq s_0$ and $\|\beta_*(\gamma_*)\|_1 \leq R$. Now if we choose the tuning parameter as $\lambda = c \sqrt{\frac{\log p}{n}}$ and the following holds

$$
c \geq 3\tau_1 R + 3\tau_2 + 2BM_2, \quad A \geq \frac{4(\tau_2 + \tau_1 R + c)}{M_1}, \quad n \geq \max \left\{ \frac{16c^2}{9M_1^2}, \frac{16\tau_1^2 s_0^2}{M_1^2}, \frac{4\kappa_0}{M_1^2} \right\} \log p,
$$
then there exists a unique sparse solution to the optimization problem (2.3) with at most \( s_0 \) nonzero coordinates satisfying that

\[
\|\hat{\beta}_{\gamma^*} - \beta_{\gamma^*}\|_\infty \leq \frac{A\sigma}{2} \sqrt{\frac{\log p}{n}} \quad \text{and} \quad \hat{\beta}_{(\gamma^*)^c} = 0.
\]

The results of Theorem 2.1 rely on the the \( \ell_1 \) norm of the weak signals \( \beta_{(\gamma^*)^c} \). Such condition might be removable when some key quantities can be quantified through other distributional properties and we thus present the following result

**Theorem 2.2.** Assume Condition 1 and 2 hold. We also assume \( d_* \geq A\sigma\sqrt{\frac{\log p}{n}} \), \( b_* \leq B\sigma\sqrt{\frac{\log p}{n}} \), \( s_* \leq s_0 \) and \( \|X_{\gamma^*}^TX_{(\gamma^*)^c}\|_\infty \leq \tau_3\lambda_n \). Now if we choose the tuning parameter as \( \lambda = c\sqrt{\frac{\log p}{n}} \) and the following holds

\[c \geq 3\tau_3 + 3\tau_2 + 2BM_2, \quad A > \frac{4(\tau_2 + \tau_3 + c)}{M_1}, \quad n \geq \max \left\{ \frac{16c^2}{9M_1^2}, \frac{16\tau_1^2 s_0^2}{M_1^2}, \frac{4\kappa_0}{M_1^2} \right\} \log p,\]

then there exists a unique sparse solution to the optimization problem (2.3) with at most \( s_0 \) nonzero coordinates satisfying that

\[
\|\hat{\beta}_{\gamma^*} - \beta_{\gamma^*}\|_\infty \leq \frac{A\sigma}{2} \sqrt{\frac{\log p}{n}} \quad \text{and} \quad \hat{\beta}_{(\gamma^*)^c} = 0.
\]

The two theorems characterize the behavior of (2.3) in three aspects: the minimum signal strength, the \( \ell_\infty \) loss on the estimation and the feature selection consistency. In particular, it is shown that there exists a unique sparse local optimizer (sparsity rate up to \( s_0 \)) that possesses a good feature selection property, regardless of the nonconvexity of (2.3). The proofs of the two theorems are provided in the appendix B.

### 2.3 Bayesian variable selection

The Bayesian variable selection approach selects the best model based on the posterior probability \( \pi_n(\gamma|X,Y) \) after placing a prior on the model space. The advantage
of such approach is the flexibility and robustness in reporting multiple models with proper uncertainty measures in contrast to the single model reported in the Frequentist approach. In this section, we introduce some background on Bayesian variable selection and the convergence theory.

2.3.1 Prior specification and sampling

The prior distribution is defined on the binary inclusion indicator variable for each feature $\gamma_i \in \{0, 1\}$. In particular, $\gamma_i = 1$ stands for that the $i^{th}$ feature is selected and vice versa. From now on and throughout the thesis, $\gamma$ will be used to denote either the binary inclusion indicator vector or the model indexes set (indexes for which $\gamma_i = 1$) interchangeably when there is no ambiguity. To fulfill the feature selection purpose, a prior distribution praising sparse models is necessary. The concrete inference in this section follows Yang et al. (2015). More precisely, the inference model is defined as follows.

**Condition 3.** The model and the prior distribution follows

$$
Y|\gamma, \beta, \phi \sim N(X_\gamma \beta_\gamma, \phi^{-1} I_n) \\
\beta_\gamma|\gamma, \phi \sim N(0, g\phi^{-1}(X_\gamma^T X_\gamma)^{-1}) \\
\pi(\phi) \sim \phi^{-1} \\
\pi(\gamma) \sim \frac{1}{p^{s_0}}I(|\gamma| \leq s_0),
$$

where $g = p^{2\alpha}$ and $\kappa$ are the two hyper parameters.

The parameter $g$ is induced from the Zellner’s $g$-prior Zellner (1986) and is usually specified as $g = p^3$ in many applications (Yang et al., 2015). $\kappa$ is the parameter controlling the prior concentration: larger $\kappa$ place more prior probability on the small models. The concrete choice of $g$, $\alpha$ and $\kappa$ will be specified in the theory part. Notice that the defined model assumes $\epsilon$ is drawn from a Gaussian family, which might not
be true for the underlying true model. However, as is shown later, the inference can still be correct even with the mis-specified model. Under the aforementioned model, the posterior distribution follows (after integrating out $\beta$ and $\phi$)
\[
\pi_n(\gamma|Y) \approx \frac{1}{p^{\|\gamma\|}} \frac{(1 + g)^{-\|\gamma\|}}{(1 + g(1 - R_\gamma^2))^{\frac{N}{2}}} f(|\gamma| \leq s_0)
\]
(2.4)
where $R_\gamma^2 = Y^T \Phi_\gamma Y/\|Y\|^2$ with the projection matrix $\Phi_\gamma = X_\gamma(X_\gamma^TX_\gamma)^{-1}X_\gamma^T$.

A direct optimization over $\pi_n(\gamma|Y)$ for searching the model with the highest posterior probability is unwise, as such optimization problem is intractable due to the complexity issue. The typical approach that Bayesian methods take is to sample from the posterior distribution via the Markov Chain Monte Carlo (MCMC) algorithm, and make inference over the posterior realizations. We refer the readers to Robert and Casella (2013) for more comprehensive review on general MCMC algorithms. Here we present a single algorithm (Algorithm 2) in Yang et al. (2015) for drawing posterior samples from (2.4).

**Algorithm 2** the MCMC algorithm

1: Input $(Y, X), s_0$ and $N$. Standardize $X$ and $Y$ to $x$ and $y$ with mean zero;
2: Initialize the model $\gamma^{(0)}$;
3: for iteration $t = 1$ to $N$ do
4: $\gamma = \gamma^{(t-1)}$;
5: Flip a coin $u \sim \text{Bern}(0.5)$;
6: if $u = 1$ then
7: Sample $j \sim \text{uniform}(1, 2, \cdots, p)$;
8: $\gamma_j = 1 - \gamma_j^{(t-1)}$;
9: else if $|\gamma_j^{(t-1)}| > 0$ and $|\gamma_k^{(t-1)}| < s_0$
10: Uniformly sample $j \sim \gamma^{(t-1)}$ and $k \sim (\gamma^{(t-1)})^c$;
11: $\gamma_j = 1 - \gamma_j^{(t-1)}$ and $\gamma_k = 1 - \gamma_k^{(t-1)}$;
12: Draw $u' \sim \text{uniform}([0, 1])$;
13: if $\log \pi_n(\gamma|Y) - \log \pi_n(\gamma|Y) > \log u'$ then
14: $\gamma^{(t)} = \gamma$;
15: else
16: $\gamma^{(t)} = \gamma^{(t-1)}$;
17: return The chain $\{\gamma^{(0)}, \gamma^{(1)}, \cdots, \gamma^{(N)}\}$;

As shown in the next section, the underlying true model possesses an overwhelm-
ing posterior probability under suitable conditions. Thus, it is guaranteed to be sampled by Algorithm 2 with high probability.

2.3.2 The non-asymptotic theory on posterior concentration

To guarantee the posterior distribution concentrate around the underlying true model, we need conditions on the feature matrix structure and norm bound of the weak signals similar to the regularized regression. Nevertheless, the conditions in this section could be more stringent than those in Section 2.2.2, because concentration is a much stronger requirement compared to just being the optimal posterior model (the model with the highest posterior probability). The latter would be sufficient if optimization over $\pi_n(\gamma|Y)$ is possible, while the former is required to ensure a good sampling performance. In particular, one need to control the magnitude of both $\frac{1}{n} \|X_{\gamma*} \beta_{\gamma*}\|_2^2$ and $\frac{1}{n} \|X_{(\gamma*)^c} \beta_{(\gamma*)^c}\|_2^2$ and thus we have the following result

**Theorem 2.3.** Assume Condition 2 and 3. We also assume that $r^t \sigma^2 \leq \frac{1}{n} \|w\|_2^2 \leq r^t \sigma^2$, $\frac{1}{n} \|X_{(\gamma*)^c} \beta_{(\gamma*)^c}\|_2^2 \leq q_1^t \sigma^2$, $\frac{1}{n} \|X_{\gamma*} \beta_{\gamma*}\|_2^2 \leq q_2^t \sigma^2$ and $\|\beta_{(\gamma*)^c}\|_1 \leq R$. If $d_* \geq A \sigma \sqrt{\frac{\log p}{n}}$ and the following hold

$$\alpha > \frac{16 \tau_1^2 + 26 R^2 \tau_1^2}{r^t M_1} + 3 - \kappa, \quad A^2 \geq \frac{32}{M_1} (4 q_1 + 7 q_2 + 4 r) (\alpha + \kappa + 5), \quad 12 q_1 + 12 r \leq q_2^t 2^\alpha$$

and

$$n \geq \max \left\{ \frac{8 \tau_2^2 s_0}{r^t M_1} , \frac{4 s_0^2 \tau_1^2 \sigma^2}{M_1^2} , \frac{24^2 \tau_1^2 R^2}{r^t 2 \sigma^2} \right\} \log p,$$

then the posterior probability on the strong signals satisfies that

$$\pi_n(\gamma_*|Y) \geq 1 - 3 p^{-2}.$$

Theorem 2.3 characterizes the posterior behavior under our prior specification. It illustrates the necessary conditions on the strong and weak signals as well as the relationship between the sample size and the dimensionality. The result presented in
Theorem 2.3 differs slightly from the result in Yang et al. (2015). Yang et al. (2015) require $\frac{1}{n} \| X_{\gamma_n^* \beta_n^*} \|_2^2$ to degenerate in a rate of $\log p/n$, while no such requirement is needed in Theorem 2.3, i.e., it can be arbitrarily large.

The proof of Theorem 2.3 relies on depicting the relationship between an arbitrary model $\gamma$ and the underlying true model $\gamma^*_n$ in the posterior distribution and a keen observation to connect such relationship with Condition 2. The concrete proof is provided in the Appendix C.

2.4 Preconditioning and the elliptical family

In Section 2.2.2 and 2.3.2, I have shown that the regularized regression and the Bayesian variable selection are both consistent under some regularity conditions. In particular, Condition 2 (Section 2.2.2) is the common condition imposed on the feature matrix that are required by both methods. However, such condition might not be satisfied in practice when the features are arbitrary. For example, $\tau_1$ could be big for highly correlated features, resulting in a bad convergence rate in Theorem 2.1 and 2.3. In this section, I introduce two powerful tools that can ameliorate this issue and enable the two algorithms applicable to broader families of data.

2.4.1 Preconditioning for high dimensional linear regression

Preconditioning is traditionally a technique used for improving convergence rate in solving large-scale linear systems

$$ AX = b, $$

when $A$ is ill-conditioned. Such solver is widely used in many applications, such as PDE equations and the Newton method. The basic idea is to multiply a non-singular matrix $M$ (preconditioner) from the left

$$ MAX = Mb, $$

16
so that the new matrix $MA$ is well-conditioned and the solution of the new system remains the same as the original problem. The regression problem is facing a similar issue when features are highly correlated. Recall the linear model

$$Y = X\beta + \varepsilon.$$ 

When features (columns of $X$) are highly correlated, matrix $X$ is ill-conditioned, adding difficulties in searching for good solutions in the feature selection problem. This motivates us to use preconditioning to reduce the condition number of $X$, which could bring at least two gains in our formulation. The first is the speed gain that is a direct consequence of preconditioning. The second is the correlation drop after preconditioning, making Condition 2 easier to satisfy in practice.

We motivate the preconditioning framework from the singular value decomposition similar to Jia and Rohe (2012) for high dimensional linear models ($p > n$). Consider the SVD on $X$ as $X = VDU^T$ where $V$ is an $n \times n$ orthogonal matrix, $D$ is an $n \times n$ diagonal matrix and $U$ is an $p \times n$ matrix belongs to the Stiefel manifold $V_{n,p}$. The simplest way to reduce the correlation between features is to multiply $X$ by $VD^{-1}V^T$ from the left. Thus, the very preconditioner I consider in this section and throughout the thesis is defined as

$$F = \sqrt{p}VD^{-1}V^T = (XX^T/p)^{-\frac{1}{2}}.$$ 

The linear model after preconditioning becomes

$$\frac{(XX^T/p)^{\frac{1}{2}}Y}{\tilde{Y}} = \sqrt{p}VU^T \beta + \frac{(XX^T/p)^{\frac{1}{2}}\varepsilon}{\tilde{\varepsilon}}, \quad (2.5)$$

We define $\tilde{Y}$ and $\tilde{X}$ as the decorrelated data after preconditioning and the linear form $\tilde{Y} = \tilde{X}\beta + \tilde{\varepsilon}$ still exists. In next section (Section 2.4.2), we show that the decorrelated data admits many good properties under the elliptical family.
2.4.2 The elliptical family

The elliptical family is widely seen in practice and includes many important distributional families, such as the Gaussian distribution, spherical t-distribution and many spatial distributions. Any distribution within the elliptical family admits the following density function

\[ p(x) \sim g(x^T \Sigma^{-1} x), \]

where \( g(\cdot) \) is a non-negative function and \( \Sigma \) is the covariance between variables. Alternatively, we might represent any elliptical distribution using a scaled Gaussian distribution. More precisely, we have the following result

**Lemma 1.** Assume a \( p \)-variate random vector \( x \) follows an elliptical distribution, then it admits an equivalent representation as

\[ x \overset{(d)}{=} L \frac{\sqrt{p} z}{\|z\|_2} \Sigma^{1/2} = \frac{\sqrt{p} L}{\|z\|_2} z \Sigma^{1/2}, \]

where \( z \) is \( p \)-variate standard Gaussian vector and \( L \) is a non-negative random variable independent of \( z \).

It is easy to show that \( L = \frac{x^T \Sigma^{-1} x}{p} \) which is closely related to the second order moment of \( x \). Using Lemma 1, we can now denote any elliptical distribution using the notion \( EN(\Sigma) \).

Assume \( X \) contains \( n \) independent realizations from \( EN(\Sigma) \). Following is a key property of \( X \) and is also one of the major contributions of this thesis. Denote the \( k^{th} \) row of \( X \) by \( x^{(k)} \)

**Theorem 2.4.** Assume \( x^{(k)} \sim EN(\Sigma), k \in [n] \) and \( p > c_0 n \). Define \( \bar{X} = (XX^T/p)^{-\frac{1}{2}} X \).

Then for any \( C > 0 \), there exists some constant \( 0 < c_1 < 1 < c_2 \) and \( c_3 > 0 \) such that for any \( i \neq j \in [p] \)

\[ \mathbb{P} \left( \frac{1}{n} \| \bar{x}_i \|_2 < \frac{c_1 c^*}{c^*} \right) \leq 2e^{-Cn}, \quad \mathbb{P} \left( \frac{1}{n} \| \bar{x}_i \|_2 > \frac{c_2 c^*}{c^*} \right) \leq 2e^{-Cn}, \]

18
and
\[ P\left( \frac{1}{n} |x_i^T \tilde{x}_j| > \frac{c_4 c^* t}{c_\star} \frac{1}{\sqrt{n}} \right) \leq 5e^{-Cn} + 2e^{-t^2/2}, \]
for any \( t > 0 \), where \( c_4 = \sqrt{\frac{c_2 (c_0 - c_1)}{c_3 (c_0 - 1)}} \) and \( c_\star, c^* \) are the smallest and largest eigenvalues of \( \Sigma \). In addition, for any \( \gamma \subseteq [p] \) and any vector \( \beta \in \mathbb{R}^p \), we have
\[ P\left( \frac{1}{n} \| \tilde{X}_\gamma \beta \|_2^2 \leq \frac{c_1 c_\star}{c^*} \| \Sigma^{1/2} \beta \|_2^2 \right) \leq 2e^{-Cn}, \quad P\left( \frac{1}{n} \| \tilde{X}_\gamma \beta \|_2^2 \geq \frac{c_2 c^*}{c_\star} \| \Sigma^{1/2} \beta \|_2^2 \right) \leq 2e^{-Cn}. \]

The proof of Theorem 2.4 relies on careful analysis on the structure of the Stiefel manifold and the concrete proof is provided in Appendix A.1. Another crucial quantity in quantifying the theoretical performance of the regularized regression and the Bayesian variable selection is \( \| X^T w \|_\infty \), for which elliptical random variables also perceive good properties. In particular we have

**Theorem 2.5 (Concentration with heavy-tailed errors).** Assume \( x^{(k)} \sim EN(\Sigma), k \in [n] \) with \( p > c_0 n \) and \( \text{var}(\varepsilon) = \sigma^2 \). Define \( \tilde{X} = (XX^T/p)^{-\frac{1}{2}}X \) and \( \tilde{\varepsilon} = (XX^T/p)^{-\frac{1}{2}}\varepsilon \).

If \( E[\tilde{X}_k^T (\Sigma^{1/2} - X^{(k)})] = E[L^{-2}] = l_2 \) and \( p/\log p \geq 16c_0/(\sqrt{c_0} - 1)^2 \), then for any \( t > 0 \) and \( r > 0 \), we have
\[ P\left( \frac{1}{n} \| \tilde{X}^T \tilde{\varepsilon} \|_\infty \geq \frac{r}{\sqrt{n}} \right) \leq C_1 \frac{t^2 \sigma^2 \log p}{r^2} + 2npe^{-\frac{r^2}{2}} + 2pe^{-Cn} + 3p^{-3}, \]
where \( C_1 = \frac{14c_2 c^* l_2^2}{v_1 c_\star^2} \) and \( c_2, c_\star, c^* \) are defined in Theorem 2.4 and \( v_1 \) defined in Lemma 4. If it is further true that \( E[\varepsilon_k^4 / L_k^4] \leq M_4 \) for some \( M_4 > 0 \), then we have
\[ P\left( \frac{1}{n} v_1 l_2 \sigma^2 \leq \frac{1}{n} \| \varepsilon \|_2^2 \leq 4l_2 v_2 \sigma^2 \right) \geq 1 - \frac{M_4}{n} - 4p^{-2}. \]

Theorem 2.5 captures the error behavior when heavy-tailed noise is presented. The theorem is surprising in that the correlation between \( \tilde{X} \) and \( \tilde{\varepsilon} \) can be roughly
upper bounded by a rate of $\sqrt{\frac{\log p}{n}}$ when both $X$ and $\varepsilon$ are heavy-tailed, which is previously known to hold only for $X$ and $\varepsilon$ being light-tailed. In fact, when both $X$ and $\varepsilon$ are light-tailed, for example, following Gaussian distribution, we might improve the probability rate in Theorem 2.5. Thus, we have the following result

**Theorem 2.6** (Concentration with light-tailed noise). Assume $x^{(k)} \sim N(0, \Sigma), k \in [n]$ with $p > c_0 n$ and $\varepsilon \sim N(0, \sigma^2)$. Define $\bar{X} = (XX^T/p)^{-\frac{1}{2}} X$ and $\bar{\varepsilon} = (XX^T/p)^{-\frac{1}{2}} \varepsilon$. Now for any $C > 0$, if $p/\log p \geq 4/v_1$, then for any $t > 0$,

$$\Pr\left( \frac{1}{n} \|X^T\bar{\varepsilon}\|_\infty \leq C_3 \frac{\sigma t}{\sqrt{n}} \right) \geq 1 - 2pe^{-t^2/2} - 2p^{-2} - 2pe^{-cn},$$

and

$$\Pr\left( \frac{1}{2c^*v_2} \leq \frac{1}{n} \|\bar{\varepsilon}\|_2^2 \leq \frac{5}{2c^*v_1} \right) \geq 1 - 2p^{-2} - 2e^{-n/4}.$$ 

where $C_3 = \sqrt{\frac{c_0^2c^*}{c_0c^*v_1}}$ and $C, c_1, c_2, c^*, c^*$ are the same constants defined in Theorem 2.4 and $v_1, v_2$ is defined in Proposition 4.

The proof for Theorem 2.5 and 2.6 are provided in the Appendix A.2. The three theorems presented in this section serve as the fundamental building bricks for the entire theoretical framework in this thesis. We will develop theoretical results based on these three theorems in all latter chapters.
In this chapter, we develop an embarrassingly parallel framework based on sample space partitioning. As entailed in Chapter 1, the typical procedure for parallelization via sample space partitioning is to partition the full data samples into several subsets, restoring on multiple machines, and then process simultaneously. Processing on subsets in parallel can lead to two types of computational gains. The first reduces time for calculations within each iteration of optimization or sampling algorithms via faster operations, of which, however, the computational gain might be limited by the increased cost of communication. This motivates the second approach, which conducts computations completely independently on the different subsets, and then combines the results to obtain the final output. This limits communication to the final combining step, and may lead to simpler and much faster algorithms. However, the second approach requires an effective aggregation rule to synthesize the subset results, so that the information or accuracy loss due to data partition is limited. It is keen to design an algorithm that can preserve (or improve) the statistical accuracy.
gained from the large scale of dataset.

The recent line of work in the second approach focuses on exploiting different aggregation methods. Mann et al. (2009); Zhang et al. (2012) considers using averaging to combine the estimators obtained in each subset. Assuming \( \hat{\beta}^{(i)}, i = 1, 2, \cdots, m \) are the \( m \) estimators obtained from the partitioned subsets, then the aggregated result is computed by

\[
\hat{\beta} = \frac{1}{m} \sum_{i=1}^{m} \hat{\beta}^{(i)}.
\]

In the simple case where \( \hat{\beta}^{(i)} \)'s are unbiased, it is easy to show that \( \hat{\beta} \) achieves the same mean squared error as if full data are used. For the general case where \( \hat{\beta} \) might be biased, Zhang et al. (2012) shows that under certain regularity condition

\[
\mathbb{E}\|\hat{\beta} - \beta_*\|_2^2 \leq O\left(\frac{1}{n} + \frac{m^2}{n^2}\right),
\]

which indicates that the combined estimator \( \hat{\beta} \) can achieve the same convergence rate as the full data estimator if \( m \) is chosen to be \( \sqrt{n} \). In a different work, Minsker et al. (2015) suggests to use the geometric median to combine the estimations. In particular, the geometric median is defined as

\[
\hat{\beta} = \min_{\beta} \sum_{i=1}^{m} \|\beta - \hat{\beta}^{(i)}\|,
\]

where \( \cdot \) is some norm defined on the Banach space. As shown in Minsker et al. (2015), the geometric median enjoys sharp convergence property and is robust to outliers.

In this chapter, I develop a simple yet effective aggregation framework for feature selection, referred to as message, based on the geometric median approach. The averaging rule is also incorporated as an optional refinement step in aiding the post-selection parallel inference. The detailed framework will be fully described in the
next section. There are related methods, which were proposed with the very different
goal of combining results from different imputed data sets in missing data contexts
Wood et al. (2008). However, these methods are primarily motivated for imputation
aggregation, do not improve computational time, and lack theoretical guarantees.
Another related approach is the bootstrap Lasso (Bolasso) Bach (2008), which runs
Lasso independently for multiple bootstrap samples, and then intersects the results
to obtain the final model. Asymptotic properties are provided under fixed number
of features \( p \) fixed and the computational burden is not improved over applying
Lasso to the full data set. Our message algorithm has strong justification in leading
to excellent convergence properties in both feature selection and prediction, while
being simple to implement and computationally highly efficient.

This chapter is organized as follows. In section 3.1, I depict the details on the
message framework. In section 3.2, I provide theoretical justifications for the new
framework and show the robustness of the new parallelized framework. Section 3.4
evaluates the performance of the new method via extensive numerical experiments.
All the proofs are provided in the Appendix.

3.2 The message algorithm

Recall the linear model with has \( n \) observations and \( p \) predictors,

\[ Y = X\beta + \varepsilon, \]

where \( Y \) is the \( n \times 1 \) response vector, \( X \) is \( n \times p \) designing matrix and \( \varepsilon \) is the
observation error which is assumed to have mean zero and variance \( \sigma^2 \). We assume
\( p > n \) throughout this chapter. The fundamental idea for communication efficient
parallelized inference is to partition the data set into \( m \) subset, each of which contains
a small portion of the data \( n/m \). Separate analysis on each subset will then be carried
out and the result will be aggregated to produce the final output.
As illustrated in Chapter 2, there are rich class of algorithms for conducting feature selection on the linear model. The *message* framework developed in this section can be incorporated with any of these feature selection algorithms. For a demonstration purpose, we specialize our attention to the two types of algorithms introduced in Chapter 2 and show how the new framework can be applied to accelerate the two algorithms for data with large sample size.

As briefly discussed in the introduction section, averaging and median possess different advantages. To carefully adapt these features to regression and classification, we propose the median selection subset aggregation (*message*) algorithm, which is motivated as follows. Averaging of sparse regression models leads to an inflated number of features having non-zero coefficients, and hence is not appropriate for model aggregation when feature selection is of interest. When conducting Bayesian variable selection, the median probability model has been recommended as selecting the single model that produces the best approximation to model-averaged predictions under some simplifying assumptions Barbieri and Berger (2004). The median probability model includes those features having inclusion probabilities greater than 1/2. We can apply this notion to subset-based inference by including features that are included in a majority of the subset-specific analyses, leading to selecting the ‘median model’. Let \( \gamma^{(i)} = (\gamma_{1}^{(i)}, \cdots, \gamma_{p}^{(i)}) \) denote the vector of feature inclusion indicators for the \( i^{th} \) subset, with \( \gamma_{j}^{(i)} = 1 \) if feature \( j \) is included so that the coefficient \( \beta_{j} \) on this feature is non-zero, with \( \gamma_{j}^{(i)} = 0 \) otherwise. The inclusion indicator vector for the median model \( M_{\gamma} \) can be obtained by

\[
\gamma = \arg \min_{\gamma \in \{0,1\}^{p}} \sum_{i=1}^{m} \| \gamma - \gamma^{(i)} \|_{1},
\]

or equivalently,

\[
\gamma_{j} = \text{median}\{\gamma_{j}^{(i)}, i = 1, 2, \cdots, m\} \text{ for } j = 1, 2, \cdots, p.
\]
If we apply the regularized regression or Bayesian variable selection to the full data set, in the presence of heavy-tailed observation errors, the estimated feature inclusion indicator vector will converge to the true inclusion vector at a polynomial rate. It is shown in the next section that the convergence rate of the inclusion vector for the median model can be improved to be exponential, leading to substantial gains in not only computational time but also feature selection performance. The intuition for this gain is that in the heavy-tailed case, a proportion of the subsets will contain outliers having a sizable influence on feature selection. By taking the median, we obtain a central model that is not so influenced by these outliers, and hence can concentrate more rapidly. As large data sets typically contain outliers and data contamination, this is a substantial practical advantage in terms of performance even putting aside the computational gain. After feature selection, we obtain estimates of the coefficients for each selected feature by averaging the coefficient estimates from each subset, following the spirit of Zhang et al. (2012). The message algorithm (described in Algorithm 3) only requires each machine to pass the feature indicators to a central computer, which (essentially instantaneously) calculates the median model, passes back the corresponding indicator vector to the individual computers, which then pass back coefficient estimates for averaging. The communication costs are negligible.

In Line 4, we impose a preconditioning step before running the regularized regression or Bayesian variable selection in order to gain better convergence rate. As shown in next section, the preconditioning can vastly aid the feature selection performance.

### 3.3 Consistency and robustness

In this section, I provide theoretical justifications for the message algorithm in terms of feature selection performance. This section consists of two parts covering both the regularized regression and the Bayesian variable selection. The key tool used in
Algorithm 3 message algorithm

Initialization:
1: Input \((Y, X), n, p, \) and \(m\);
2: partition \((Y, X)\) row-wisely into \(m\) subsets \((Y_i, X_i)\) and distribute on \(m\) machines;

Stage 1:
3: for \(i = 1\) to \(m\) do
4: Obtain \(\hat{\gamma}^{(i)}\) using either Algorithm 1 or 2 after preconditioning (Section 2.4.1);
5: Gather all subset models \(\hat{\gamma}^{(i)}\) to obtain the median model \(\hat{\gamma}\);
6: for \(j = 1\) to \(p\) do
7: \(\hat{\gamma}_j = \text{median}\{\hat{\gamma}_j^{(i)}; i = 1, 2, \cdots, m\};\)

Stage 2:
8: Redistribute the estimated model \(M_\hat{\gamma}\) to all subsets
9: for \(i = 1\) to \(m\) do
10: Estimate \(\hat{\beta}^{(i)}\) under \(\hat{\gamma}\) using Algorithm 1 or Bayesian sampling;
11: Gather all subset estimations \(\hat{\beta}^{(i)}\);
12: \(\bar{\hat{\beta}} = \frac{1}{m} \sum_{i=1}^{m} \hat{\beta}^{(i)};\)
13: return \(\bar{\hat{\beta}}, \hat{\gamma}\)

this section is the following Chernoff’s bound.

Lemma 2 (Chernoff bound). Let \(q_i, i = 1, 2, \cdots, n\) be independent Bernoulli random variables with \(\mathbb{P}(q_i = 1) = p > 1/2\). Define \(q = \sum_{i=1}^{n} q_i\). Then we have

\[
\mathbb{P}\left(q > \frac{n}{2}\right) \geq 1 - \exp\left\{- \frac{n}{2p} \left( p - \frac{1}{2} \right)^2 \right\}.
\]

The Chernoff’s bound ensures the robustness of the median operation. In particular, the uniformly slow convergence probability \(p > 1/2\) is converted to an exponential concentration for the median. The excellent theoretical property of the message algorithm inherits from this nice property.

3.3.1 Convergence rate for regularized regression

To capture the convergence rate for message using the regularized regression, we need first to quantify the convergence rate on a single subset with a sample size of \(n\). Assume \(\gamma_s\) is the strong signal set where \(\min_{i \in \gamma_s} \lvert \beta_{si} \rvert \geq d_s\) and \(\min_{i \notin \gamma_s} \lvert \beta_{si} \rvert \leq b_s\). We state the result for both Gaussian noise and the heavy-tailed noise.
Theorem 3.1 (Convergence on single subset with Gaussian noise). Assume $X \sim N(0, \Sigma)$, $\varepsilon \sim N(0, \sigma^2)$ and the preconditioning is used. We also assume $p > c_0 n$, $b_* \leq B \sigma \sqrt{\frac{\log p}{n}}$ and $\|\beta_{*(\gamma_*)}\|_1 \leq R$. Now choosing $\lambda = c \sigma \sqrt{\frac{\log p}{n}}$ and there exist some absolute constants $V_1, V_2, V_3$ and $C_3$ such that if the following hold,

$$c \geq 3 \sigma^{-1} V_1 R + 6 C_3 + 2 V_2 B, \quad d_* \geq \left(8 C_3 V_3 + 12 \sigma^{-1} V_1 V_3 R + 4 c V_3 \right) \sigma \sqrt{\frac{\log p}{n}}$$

and

$$n \geq \max \left\{ \frac{16}{9} c^2 V_3^2, \frac{12^2 V_1^2 V_3^2 \sigma_0^2}{\sigma^2}, \frac{4 \kappa_0 V_3^2}{2} \right\} \log p,$$

then with probability at least $1 - 12p^{-1}$, the solution of Algorithm 1 satisfies that

$$\|\hat{\beta}_{\gamma_*} - \beta_{\gamma_*}\|_\infty \leq \left(4 C_3 V_3 \sigma + 6 V_1 V_3 R + 2 V_3 c \sigma \right) \sqrt{\frac{\log p}{n}} \text{ and } \hat{\beta}_{(\gamma_*)} = 0,$$

The constants are defined as $V_1 = \frac{c_4 c_*}{c_4}$, $V_2 = \frac{c_2 c_*}{c_*}$, $V_3 = \frac{c_3}{c_1 c_*}$ and $c_1, c_2, c_4, C_3$ are constants defined in Theorem 2.4 and 2.6 and $c_*, c^*$ are the smallest and the largest eigenvalues of $\Sigma$.

Proof. We make use of Theorem 2.1, 2.4 and 2.6. Taking $t = 2 \sqrt{\log p}$ in Theorem 2.6 indicates that

$$\mathbb{P} \left( \frac{1}{n} \|\hat{X}^T \tilde{\varepsilon}\|_\infty \geq 2 C_3 \sigma \sqrt{\frac{\log p}{n}} \right) < 2 p e^{-C n} + 4 p^{-1},$$

which means that we can take $\tau_2 = 2 C_3$ in Condition 2. In addition, taking $t = \sqrt{6 \log p}$ in Theorem 2.4 gives

$$\mathbb{P} \left( \max_{i \neq j} \frac{1}{n} |\hat{x}_i^T \tilde{x}_j| > \frac{c_4 c_* \sqrt{6 \log p}}{c_*} \right) \leq 5 p^2 e^{-C n} + 2 p^{-1},$$

which means that $\tau_1$ can be taken as $\frac{3 c_4 c_*}{\sigma c_*}$ along with $M_1 = \frac{c_4 c_*}{c_*}$ and $M_2 = \frac{c_2 c_*}{c_*}$. Consequently, applying Theorem 2.1 with these values and $C = 3$ we have the following
result. If \( c \geq \frac{3c_{c}e^{R}}{c_{c}} + 6C_{3} + \frac{2c_{c}B}{c_{c}} \) and \( n \geq \max \left\{ \frac{16c_{c}^{2}c^{2}}{9c_{c}^{2}}, \frac{12c_{c}^{2}c^{2}}{c_{c}^{2}}, \frac{4c_{c}^{2}}{c_{c}^{2}} \right\} \log p \) and
\[
d_{*} \geq \left( \frac{8C_{3}c_{c}}{c_{c}} + \frac{12c_{c}^{2}e^{R}}{c_{c}} + \frac{4c_{c}^{2}}{c_{c}} \right) \sigma \sqrt{\log \frac{p}{n}},
\]
then with probability at least \( 1 - 6p^{2}e^{-Cn} \), we have
\[
\| \hat{\beta}_{\gamma*} - \beta_{\gamma*} \|_{\infty} \leq \left( \frac{8C_{3}c_{c}}{2c_{c}} + \frac{12c_{c}^{2}e^{R}}{2c_{c}^{2}} + \frac{4c_{c}^{2}}{2c_{c}} \right) \sqrt{\frac{\log p}{n}} \quad \text{and} \quad \hat{\beta}_{(\gamma*)c} = 0.
\]

Using the definition of \( V_{1}, V_{2}, V_{3} \) we can retrieve the results stated in the theorem.

Taking \( C = 3 \) and notice that \( n > \log p \) we achieve the \( 1 - 12p^{-1} \) convergence probability.

For \( \epsilon \) following heavy tailed distribution, we expect to see a slower convergence rate for consistent feature selection. In particular, the strong signals are required to have slightly stronger strength in order to be consistently selected.

**Theorem 3.2** (Convergence on single subset with heavy-tailed noise). Assume \( X \sim EN(\Sigma), \text{var}(\varepsilon) = \sigma^{2}, \mathbb{E}[\varepsilon^{-2}] = l_{2} \) and the preconditioning is used. We also assume \( p > c_{0}n, b_{*} \leq B\sigma \sqrt{\frac{\log p}{n}} \) and \( \| \beta_{*}(\gamma*) \|_{1} \leq R \). Now choosing \( \lambda = c\sigma \sqrt{\frac{\log p}{n}} \) and there exist some absolute constants \( V_{1}, V_{2}, V_{3} \) and \( C_{1} \) such that if the following hold,
\[
c \geq 3\sigma^{-1}V_{1}R + 3\sqrt{C_{1}n^{\frac{1}{2}}} + 2V_{2}B, \quad d_{*} \geq \left( 4\sqrt{V_{1}n^{\frac{1}{2}}} + 12\sigma^{-1}V_{1}V_{3}R + 4cV_{3} \right) \sigma \sqrt{\frac{\log p}{n}}
\]
and
\[
n \geq \max \left\{ \frac{16}{9}c^{2}V_{3}^{2}, \frac{12^{2}V_{1}^{2}V_{3}^{2}}{\sigma^{2}}, \frac{4c_{c}^{2}V_{3}^{2}}{\sigma^{2}}, 4\kappa_{0}V_{3}^{2} \right\} \log p,
\]
where \( \delta \in (0, 1) \) is an arbitrary constant, then with probability at least \( 1 - 6p^{2}e^{-Cn} \), the solution of Algorithm 1 satisfies that
\[
\| \hat{\beta}_{\gamma*} - \beta_{\gamma*} \|_{\infty} \leq 2\sqrt{C_{1}V_{3}^{2}} \sigma \sqrt{\frac{\log p}{n}} + \left( 6V_{1}V_{3}R + 2V_{3}c \right) \sigma \sqrt{\frac{\log p}{n}} \quad \text{and} \quad \hat{\beta}_{(\gamma*)c} = 0,
\]
The constants are defined as $V_1 = \frac{c_{1*}}{c_*}, V_2 = \frac{c_{2*}}{c_*}, V_3 = \frac{c_*}{c_1c_*}$ and $c_1, c_2, c_4, C_1$ are constants defined in Theorem 2.4 and 2.5 and $c_*, c^*$ are the smallest and the largest eigenvalues of $\Sigma$.

Proof. We make use of Theorem 2.1, 2.4 and 2.5. Taking $r = \sigma \sqrt{C_1 n^\delta \log p}$ and $t = \sqrt{6 \log p}$ in Theorem 2.5, we have

$$P \left( \frac{1}{n} \| \tilde{X}^T \tilde{\varepsilon} \|_\infty \geq \sqrt{C_1 \sigma \sqrt{\frac{\log p}{n^{1-\delta}}}} \right) \leq \frac{6 \log p}{n^\delta} + 5p^{-\delta} + 2pe^{-c_1n},$$

which means that we can take $\tau_2 = \sqrt{C_1 n^\delta}$ in Condition 2. In addition, taking $t = \sqrt{6 \log p}$ in Theorem 2.4 gives

$$P \left( \max_{i \neq j} \frac{1}{n} | \tilde{x}_i^T \tilde{x}_j | > \frac{c_4 c^* \sqrt{6 \log p}}{c_*} \frac{1}{\sqrt{n}} \right) \leq 5p^2 e^{-c_1n} + 2p^{-1},$$

which means that $\tau_1$ can be taken as $\frac{3c_{1*} e^*}{\sigma c_*}$ along with $M_1 = \frac{c_{1*}}{c_*}$ and $M_2 = \frac{c_{2*}}{c_*}$. Consequently, applying Theorem 2.1 with these values and $C = 3$ we have the following result. If $c \geq \frac{3c_{1*} e^* R}{\sigma c_*} + 3\sqrt{C_1 n^\delta} + 2c_{2*} \frac{B}{c_*}$ and $n \geq \max \left\{ \frac{16c_2^2 c_*^2}{9c_1^2 c_*^2}, \frac{12c_1^2 c_*^4}{c_*^2 + c_1^2}, \frac{4c_1 c_*^2}{c^2} \right\} \log p$

and $d_* \geq \left( \frac{4\sqrt{C_1 n^\delta c^*}}{c_1 c_*} + \frac{12c_4 c_*^2 R}{c_1 c_*^2} + \frac{4c_*}{c_1 c_*} \right) \sigma \sqrt{\frac{\log p}{n}},$ then with probability at least $1 - 6\frac{\log p}{n^\delta} - 5p^2 e^{-c_1n} - 7p^{-1}$ we have

$$\| \hat{\beta}_{\gamma_*} - \beta_{\gamma_*} \|_\infty \leq \left( \frac{4\sqrt{C_1 n^\delta \sigma c^*}}{2c_1 c_*} + \frac{12c_4 c_*^2 R}{2c_1 c_*^2} + \frac{4c_* \sigma}{2c_1 c_*} \right) \sqrt{\frac{\log p}{n}}$$

and $\hat{\beta}_{(\gamma_*^c)} = 0$.

Taking $C = 3$ and notice that $n > \log p$ we achieve the $1 - 12p^{-1}$ convergence probability.

It can be seen from Theorem 3.2 that the best dimensionality the regularized regression can achieve for $\| \cdot \|_\infty$ consistency is $\log p = O(\sqrt{n})$ when heavy-tailed noise is presented. However, if only cares about feature selection consistency and the signal
threshold $d_*$ is sufficiently large (for example, $O(\sqrt{\log p})$), then by choosing $\delta = 0$ one can still achieve the $\log p = O(n)$ dimensionality but the value of $\hat{\beta}_{*\gamma_*}$ could be badly deviated from $\beta_{*\gamma_*}$.

With the result on each single subset, we are now ready to state the main theorem for the message algorithm. In particular, we have

**Theorem 3.3** (message for regularized regression). *Assume the conditions in Theorem 3.1 for Gaussian noise or Theorem 3.2 for heavy-tailed noise and the total sample size is $N < p$. For each subset we choose $\lambda = c\sigma\sqrt{\log \frac{p}{n_0}}$ where $n_0$ is the subset size. If $c$ and $n_0$ satisfies the condition in Theorem 3.1 and $n_0 \geq 48$, then there exists some absolute constants $K_1 > 0$ such that

$$ \mathbb{P}(\hat{\gamma} = \gamma_*) \geq 1 - \exp \left\{ - \frac{K_1 d_*^2}{32(1 + R^2 + B^2)\sigma^2 \log p} N \right\}. $$

If $c$ and $n_0$ satisfies the condition in Theorem 3.2 and $n_0 \geq 328(\log p)^2$, then there exists some absolute constants $K_2 > 0$ such that

$$ \mathbb{P}(\hat{\gamma} = \gamma_*) \geq 1 - \exp \left\{ - \frac{K_2 d_*^4}{128(1 + R^4 + B^4)\sigma^4 (\log p)^2} N \right\}. $$

Notice the original convergence rate for heavy-tailed noise is at most $1 - O\left(\frac{\log p}{n}\right)$ regardless of the sample sizes. However, by using message, when the sample size $N > O\{(\log p)^3\}$, one can achieve a convergence rate at $1 - O(p^{-1})$ or even exponential convergence. The sample size of each subset just need to grow proportional to $(\log p)^2$.

**3.3.2 Convergence rate for Bayesian variable selection**

Similar as the previous section, we need first to capture the performance of Bayesian variable selection on each single subset. We have the following results for Gaussian and Heavy-tailed noise
Theorem 3.4 (Convergence on single subset with Gaussian noise). Assume \( X \sim N(0, \Sigma) \), \( \varepsilon \sim N(0, \sigma^2) \) and the preconditioning is used. If \( p > c_0 n \), \( b_* \leq B \sigma \sqrt{\log p \over n} \) and \( \| \beta_{s(\gamma^c)} \|_1 \leq R \), then there exist some constants \( V'_1, V'_2, V'_3, V'_4 \) and \( V'_5 \) such that if the following conditions hold

\[
\alpha \geq V'_1 + 9V'_2V'_4\sigma^{-2}R^2 + 3 - \kappa, \quad {d_s^2 \over \alpha + \kappa + 5} \geq V'_3 \left( {4BR \over K_0} \sqrt{\log p \over n} + V'_5 + 7 \right) \sigma^2 \log p \over n
\]

and

\[
\| \beta_{s \gamma} \|_2^2 \leq K_0 p^{2\alpha}, \quad n \geq \max \left\{ {1 \over 2} V'_1 s_0, V'_2 s_0^2, 72\sigma^{-2}V'_4 V'_4R^2 \right\} \log p
\]

then with probability at least \( 1 - 13 p^{-1} - 2e^{-n/4} \),

\[
\pi_n (\gamma_* | Y) \geq 1 - 3p^{-2},
\]

for \( p \geq \max \{ (12BR\sqrt{\log p \over n} + 3V'_5)^{(2\alpha - 1)} / 5 \} \). The constants are defined as \( V'_1 = {128C_3^2c^2v_2 \over c_1c_*} \), \( V'_2 = {36c^2^4 \over c_1^4c_*} \), \( V'_3 = {64c^4v_2K_0 \over c^2_*} \), \( V'_4 = c_1c_*v_2 \) and \( V'_5 = {10 \over c_1c^2v_2K_0} \) where \( c_1, c_2, c_4, c_*, c^* \), \( v_1, C_3 \) and \( v_2 \) are defined in Theorem 2.4, 2.6 and Proposition 4.

Proof. We just need to verify all conditions for Theorem 2.3. The verification of Condition 2 is similar to Theorem 3.1. To bound the additional norm constraints on \( \tilde{\varepsilon}, \tilde{X}_\gamma, \beta_\gamma \) and \( \tilde{X}_{(\gamma)^c}, \beta_{(\gamma)^c} \), we use Theorem 2.6 so that

\[
\mathbb{P} \left( \frac{1}{2c^* v_2} \leq \frac{1}{n} \| \tilde{\varepsilon} \|_2^2 \leq \frac{5}{2c_4 v_1} \right) \geq 1 - 2p^{-2} - 2e^{-n/4},
\]

which indicates that \( r' = {1 \over 2c^* v_2} \) and \( r = {5 \over 2c_4 v_1} \). In addition, using the second part of Theorem 2.4 for \( \gamma = \gamma_* \) and \( \gamma = (\gamma_*)^c \) we have with probability at least \( 1 - 4e^{-Cn} \)

\[
q_2 = {c_1c^2K_0 \over c_*} \quad \text{and} \quad q_1 = {c_1c^2BR \sigma \over c_*} \sqrt{\log p \over n}.
\]
From the proof of Theorem 3.1, we know that \( \tau_2 = 2C_3 \) with probability at least \( 1 - 2pe^{-Cn} - 4p^{-1} \), \( \tau_1 = \frac{3c_4c^*}{\sigma c^*} \) with probability at least \( 1 - 5p^2e^{-Cn} - 2p^{-1} \) and \( M_1 = \frac{c_4c^*}{c^*}, M_2 = \frac{c_4c^*}{c^*} \) with probability at least \( 1 - 4pe^{-Cn} \). By summarizing all results and take \( C = 3 \) we obtain that if

\[
\alpha \geq \frac{128C_3^2c^* v_2}{c_1 c^*} + \frac{234c_4^2c^*^4v_2R^2}{c_1 c^*^2\sigma^2} + 2 - \kappa
\]

and

\[
\frac{A^2}{\alpha + \kappa + 5} \geq \frac{64c^* v_2K_0}{c_2^*} \left( \frac{4BR}{K_0} \sqrt{\frac{\log p}{n}} + 7 + \frac{10}{c_1 c^* v_1 K_0} \right)
\]

and

\[
n \geq \max \left\{ \frac{64C_3^2 c^* v_2 s_0}{c_1 c^*}, \frac{36c_4^2 c^*^4 s_0^2}{c_2^* c^*^4}, \frac{9 \times 2^{10} C_3^2 c^* v_2^2 R^2}{\sigma^2} \right\} \log p
\]

and

\[
p^{2\alpha} \geq \frac{12BR}{K_0} \sqrt{\frac{\log p}{n}} + \frac{30}{c_1 c^* v_1 K_0},
\]

then for \( p \geq 5 \) with probability at least \( 1 - 13p^{-1} - 2e^{-n/4} \),

\[
\pi_n(\gamma_*|Y) \geq 1 - 3p^{-2}.
\]

\( \square \)

**Theorem 3.5** (Convergence on single subset with heavy-tailed noise). Assume \( x^{(k)} \sim \mathcal{N}(L_k, \Sigma) \), \( \text{var}(\varepsilon_k) = \sigma^2 \), \( \mathbb{E}[L_k^{-2}] = l_2 \) and \( \mathbb{E}[\varepsilon_k^4/L_k] \leq M_4, k \in [n] \). We also assume the preconditioning is used. If \( p > c_0n, b_* \leq B\sigma\sqrt{\frac{\log p}{n}} \) and \( \|\beta_*(\gamma_*)\cdot\|_1 \leq R \), then there exist some constants \( V_1', V_2', V_3', V_4' \) and \( V_5' \) such that if the following conditions hold

\[
\alpha \geq V_1' n^\delta + 18V_4' V_2' \sigma^{-2} R^2 + 3 - \kappa, \quad \frac{d_*^2}{\alpha + \kappa + 5} \geq V_3' \left( \frac{4BR}{K_0} \sqrt{\frac{\log p}{n}} + V_5' + 7 \right) \frac{\sigma^2 \log p}{n}
\]

32
and
\[ \| \beta_n \|^2 \leq K_0 p^{2\alpha}, \quad n \geq \max \left\{ \frac{1}{2} V_1'' s_0 n^\delta, V_2' s_0^2, 144\sigma^{-2} V_4'' R^2 n^\delta \right\} \log p \]

where \( \delta \in (0, 1) \) is an arbitrary constant, then with probability at least \( 1 - 6 \log p/n^\delta - M_4/n - 15p^{-1} \),
\[ \pi_n(\gamma_*|Y) \geq 1 - 3p^{-2}, \]

for \( p \geq \max\{ (\frac{12B R \sqrt{\log p}}{K_0 \sqrt{n}})^{1/(2\alpha)} + 3V_5'' \}^{1/(2\alpha)}, 5, 4v_1^{-1} \log p \} \). The constants are defined as
\[ V_1'' = \frac{64C_1 c_*}{c_1 c_* v_1 l_1}, V_2' = \frac{36c_1^2 c_*^4}{c_1^4 c_*^4}, V_3' = \frac{64c_*^4 v_2 K_0}{c_*^4}, V_4'' = \frac{c_1 C_*}{c_*^2 v_1 l_1}, \]
and \( V_5'' = \frac{16c_* v_2 l_2}{c_1^2 c_*^2 K_0} \) where \( c_1, c_2, c_4, c_*, c, v_1, C_3 \) and \( v_2 \) are defined in Theorem 2.4, 2.6 and Proposition 4.

**Proof.** Similar to the proof of Theorem 3.4, we just need to quantify the norms under the heavy tailed noise. Using Theorem 2.5, we know that \( r' = \frac{1}{4} v_1 l_1 \) and \( r = 4v_2 l_2 \) with probability at least \( 1 - M_4/n - 4p^{-2} \). The value of \( \tau_2 \) follows the same argument in the proof of Theorem 3.2, i.e., with probability at least \( 1 - 6 \log p/n^\delta - 5p^{-1} - 2pe^{-Cn} \) we can choose \( \tau_2 = \sqrt{C_1 n^\delta} \). The rest quantities can be bounded the same as in Theorem 3.4. Thus taking \( C = 3 \) we achieve the following result: if

\[ \alpha \geq \frac{64C_1 c_*}{c_1 c_* v_1 l_1} n^\delta + \frac{468c_4^2 c_*^3 R^2}{c_1^2 c_*^2 v_1 l_1 \sigma^2} + 3 - \kappa \]

and

\[ \frac{A^2}{\alpha + \kappa + 5} \geq \frac{64c_*^4 v_2 K_0}{c_*^2} \left( \frac{4BR}{K_0} \sqrt{\frac{\log p}{n}} + 7 + \frac{16c_* v_2 l_2}{c_1^2 c_*^2 K_0} \right) \]

and

\[ n \geq \max \left\{ \frac{32C_1 c_* s_0}{c_1 c_* v_1 l_1} n^\delta, \frac{36c_4^2 c_*^4 s_0^2}{c_1^2 c_*^4}, \frac{9 \times 2^{10} C_1 R^2}{v_1^2 l_1^2 \sigma^2} - n^\delta \right\} \log p \]

and

\[ p^{2\alpha} \geq \frac{12B R \sqrt{\log p}}{K_0 \sqrt{n}} + \frac{48C_* v_2 l_2}{c_1 c_*^2 K_0}; \]
then for \( p \geq 5 \) with probability at least \( 1 - 6 \log p / n^\delta - M_4 / n - 15p^{-1} \),

\[
\pi_n(\gamma_*|Y) \geq 1 - 3p^{-2}.
\]

For the heavy-tailed noise, the Bayesian method requires a fourth-moment condition on the distribution, which is slightly stronger condition compared to the Frequentist method. This is a direct consequence of the fact that posterior concentration is a stronger requirement than the true model having the largest posterior probability. For Theorem 3.1 and 3.5 we proved the result for a concentration rate at \( 1 - O(p^{-2}) \), which is used to ensure a concentration rate at \( 1 - O(p^{-1}) \) uniformly for all subsets. This allows us to leverage the model with highest posterior probability via message. We now proceed to state the main result for message in using Bayesian methods.

**Theorem 3.6 (message for Bayesian variable selection).** Assume the conditions in Theorem 3.4 for Gaussian noise or Theorem 3.5 for heavy-tailed noise and the total sample size is \( N < p \). For each subset we allocate \( n_0 \) samples and report the model within the sampled models that has the highest posterior probability. If \( \alpha, \kappa \) and \( n_0 \) satisfies the condition in Theorem 3.4 and also \( n_0 \geq 60 \), then there exists some absolute constants \( K'_1 > 0 \) such that

\[
\Pr(\hat{\gamma} = \gamma_*) \geq 1 - \exp \left\{ - \frac{K'_1d^2_s}{32(1 + R^2 + RB)\sigma^2 \log p} N \right\}.
\]

If \( \alpha, \kappa \) and \( n_0 \) satisfies the condition in Theorem 3.5 and \( n_0 \geq 328(\log p)^2 \), then there exists some absolute constants \( K'_2 > 0 \) such that

\[
\Pr(\hat{\gamma} = \gamma_*) \geq 1 - \exp \left\{ - \frac{K'_2d^2_s}{128(1 + R^2 + RB)\sigma^2 (\log p)^2} N \right\}.
\]
3.4 Experiments

We use lasso as an example to illustrate the performance of message. This section assesses the performance of the message algorithm via extensive examples, comparing the results to

- Full data inference. (denoted as “full data”)

- Subset averaging. Partition and average the estimates obtained on all subsets. (denoted as “averaging”)

- Subset median. Partition and take the marginal median of the estimates obtained on all subsets (denoted as “median”)

- Bolasso. Run Lasso on multiple bootstrap samples and intersect to select model. Then estimate the coefficients based on the selected model. (denoted as “Bolasso”)

The Lasso part of all algorithms will be implemented by the “glmnet” package Friedman et al. (2010). (We did not use ADMM Boyd et al. (2011) for Lasso as its actual performance might suffer from certain drawbacks Peng et al. (2013) and is reported to be slower than “glmnet” Li et al. (2013))

3.4.1 Synthetic data sets

We use the linear model and the logistic model for \((p; s) = (1000; 3)\) or \((10,000; 3)\) with different sample size \(n\) and different partition number \(m\) to evaluate the performance. The feature vector is drawn from a multivariate normal distribution with correlation \(\rho = 0\) or 0.5. Coefficients \(\beta\) are chosen as,

\[
\beta_i \sim (-1)^{\text{ber}(0.4)}(8 \log n / \sqrt{n} + |N(0, 1)|), i \in S
\]

We utilize lasso for evaluating the performance of message. The concrete setup of models are as follows,
1. Linear model with $\epsilon \sim N(0, 2^2)$.

2. Linear model with $\epsilon \sim t(0, df = 3)$.

3. Logistic model.

For $p = 1,000$, we simulate 200 data sets for each case, and vary the sample size from 2000 to 10,000. For each case, the subset size is fixed to 400, so the number of subsets will be changing from 5 to 25. In the experiment, we record the mean square error for $\hat{\beta}$, probability of selecting the true model and computational time, and plot them in Fig 3.1 - 3.6. For $p = 10,000$, we simulate 50 data sets for each case, and let the sample size range from 20,000 to 50,000 with subset size fixed to 2000.

![Graphs showing mean square error, probability to select the true model, and computational time for different sample sizes and subset sizes.](image)

**Figure 3.1:** Results for case 1 with $\rho = 0$. Top: $p = 1000$; Bottom: $p = 10,000$

It is clear that `message` had excellent performance in all of the simulation cases, with low MSE, high probability of selecting the true model, and low computational time. The other subset-based methods we considered had similar computational times and also had computational burdens that effectively did not increase with
Figure 3.2: Results for case 1 with $\rho = 0.5$. Top: $p = 1000$; Bottom: $p = 10,000$

Figure 3.3: Results for case 2 with $\rho = 0$. Top: $p = 1000$; Bottom: $p = 10,000$
Figure 3.4: Results for case 2 with $\rho = 0.5$. Top: $p = 1000$; Bottom: $p = 10,000$

Figure 3.5: Results for case 3 with $\rho = 0$. Top: $p = 1000$; Bottom: $p = 10,000$
Figure 3.6: Results for case 3 with $\rho = 0.5$. Top: $p = 1000$; Bottom: $p = 10,000$

sample size, while the full data analysis and bootstrap Lasso approach both were substantially slower than the subset methods, with the gap increasing linearly in sample size. In terms of MSE, the averaging and median approaches both had dramatically worse performance than message in every case, while bootstrap Lasso was competitive (MSEs were same order of magnitude with message ranging from effectively identical to having a small but significant advantage), with both message and bootstrap Lasso clearly outperforming the full data approach. In terms of feature selection performance, averaging had by far the worst performance, followed by the full data approach, which was substantially worse than bootstrap Lasso, median and message, with no clear winner among these three methods. Overall message clearly had by far the best combination of low MSE, accurate model selection and fast computation.
3.4.2 Individual household electric power consumption

This data set contains measurements of electric power consumption for every household with a one-minute sampling rate Bache and Lichman (2013). The data have been collected over a period of almost 4 years and contain 2,075,259 measurements. There are 8 predictors, which are converted to 74 predictors due to re-coding of the categorical variables (date and time). We use the first 2,000,000 samples as the training set and the remaining 75,259 for testing the prediction accuracy. The data are partitioned into 200 subsets for parallel inference. We plot the prediction accuracy (mean square error for test samples) against time for full data, message, averaging and median method in Fig 3.7. Bolasso is excluded as it did not produce meaningful results within the time span.

To illustrate details of the performance, we split the time line into two parts: the early stage shows how all algorithms adapt to a low prediction error and a later stage captures more subtle performance of faster algorithms (full set inference excluded due to the scale). It can be seen that message dominates other algorithms in both speed and accuracy.

![Figure 3.7: Results for power consumption data.](image)
3.4.3 **HIGGS classification**

The HIGGS data have been produced using Monte Carlo simulations from a particle physics model Baldi et al. (2014). They contain 27 predictors that are of interest to physicists wanting to distinguish between two classes of particles. The sample size is 11,000,000. We use the first 10,000,000 samples for training a logistic model and the rest to test the classification accuracy. The training set is partitioned into 1,000 subsets for parallel inference. The classification accuracy (probability of correctly predicting the class of test samples) against computational time is plotted in Fig 3.8 (Bolasso excluded for the same reason as above).

![Mean prediction accuracy](image)

**Figure 3.8:** Results for HIGGS classification.

*Message* adapts to the prediction bound quickly. Although the classification results are not as good as the benchmarks listed in Baldi et al. (2014) (due to the choice of a simple parametric logistic model), our new algorithm achieves the best performance subject to the constraints of the model class.
4

Feature space partition using DECO

4.1 Introduction

In modern science and technology applications, it has become routine to collect complex datasets with a huge number $p$ of variables and/or enormous sample size $n$. However, like Chapter 3, most of the emphasis in the literature has been on addressing large $n$ problems, with a common strategy relying on partitioning data samples into subsets and fitting a model containing all the variables to each subset (Mcdonald et al., 2009; Zhang et al., 2012; Wang and Dunson, 2013; Scott et al., 2013; Wang et al., 2015c, 2014; Minsker et al., 2015). In scientific applications, it is much more common to have huge $p$ small $n$ data sets. In such cases, a sensible strategy is to break the features into groups, fit a model separately to each group, and combine the results. We refer to this strategy as feature space partitioning.

There are several recent attempts on parallel variable selection by partitioning the feature space. Song and Liang (2014) proposed a Bayesian split-and-merge (SAM) approach in which variables are first partitioned into subsets and then screened over each subset. A variable selection procedure is then performed on the variables that
survive for selecting the final model. One caveat for this approach is that the algorithm cannot guarantee the efficiency of screening, i.e., the screening step taken on each subset might select a large number of unimportant but correlated variables (Song and Liang, 2014), so the split-and-merge procedure could be ineffective in reducing the model dimension. Inspired by a group test, Zhou et al. (2014) proposed a parallel feature selection algorithm by repeatedly fitting partial models on a set of re-sampled features, and then aggregating the residuals to form scores for each feature. This approach is generic and efficient, but the performance relies on a strong condition that is almost equivalent to an independence assumption on the design.

As discussed in Chapter 1, feature space partitioning is much more challenging than sample space partitioning due to the correlations between the features. A naive partition of the feature space would succeed only when the features across the partitioned subsets were mutually independent or uncorrelated. Otherwise, it is highly likely that any model posed on the subsets is misspecified and the results are biased regardless of the sample size. As a result, some novel "decorrelation" technique is needed in shrinking the correlation between features, so that the feature space partitioning is possible. In this chapter, we propose a novel embarrassingly-parallel framework named \textit{DECO} by \textit{decorrelating} the features before partitioning. With the aid of decorrelation, each subset of data after feature partitioning can now produce consistent estimates even though the model on each subset is intrinsically mis-specified due to missing features. To the best of our knowledge, \textit{DECO} is the first embarrassingly parallel framework specifically designed to accommodate arbitrary correlation structure in the features. We show, quite surprisingly, that the \textit{DECO} estimate, by leveraging the estimates from subsets, achieves the same convergence rate in $\ell_\infty$ norm as the estimate obtained by using the full dataset, and that the rate does not depend on the number of partitions.

There are some related work that are motivated from a completely different goal.
The most related one is Jia and Rohe (2012), where a similar procedure was introduced to improve lasso. DECO differs substantially in various aspects. First, the motivation for DECO is to develop a parallel computing framework for fitting large-\(p\) data by splitting features, which can potentially accommodate any penalized regression methods, while Jia and Rohe (2012) aim solely at complying with the irrepresentable condition for lasso. Second, the conditions posed on the feature matrix are more flexible in DECO, and our theory, applicable for not only sparse signals but also those in \(l_r\) balls, can be readily applied to the preconditioned lasso in Jia and Rohe (2012).

The rest of this chapter is organized as follows. In Section 4.2, we detail the proposed framework. Section 4.3 provides the theory of DECO. Section 4.4 presents extensive simulation studies to illustrate the performance of our framework. All the technical details are relegated to the Appendix.

4.2 The DECO framework

We motivate the DECO framework based on the linear model. Recall our working model

\[
Y = X\beta + \varepsilon, \quad (4.1)
\]

where \(X\) is an \(n \times p\) feature (design) matrix, \(\varepsilon\) consists of \(n\) i.i.d random errors and \(Y\) is the response vector. Despite the large class of feature selection algorithms, as in Chapter 3, we specialize our discussion to the regularized regression and the Bayesian variable selection to highlight the main message of this section.

As discussed in the introduction, a naive partition of the feature space will usually give unsatisfactory results under a parallel computing framework, which motivates the decorrelation step. The most intuitive way to decorrelate data is to orthogonalize features via the singular value decomposition (SVD) of the feature matrix. We use
the low dimensional regression as a motivating example.

Motivating from the low dimensional model When \( n > p \), we consider the SVD of the feature matrix \( X \) with \( X = V D U^T \), where \( V \) is an \( n \times p \) matrix, \( D \) is an \( p \times p \) diagonal matrix and \( U \) is an \( p \times p \) orthogonal matrix. If we pre-multiply both sides of (4.1) by \( \sqrt{p} V D^{-1} V^T \), we get

\[
\sqrt{p} V D^{-1} V^T Y = \sqrt{p} V U^T \beta + \sqrt{p} V D^{-1} V^T \varepsilon.
\]

It is obvious that the new features (the columns of \( \sqrt{p} V U^T \)) are mutually orthogonal. Define the new data after decorrelation as \( (\tilde{Y}, \tilde{X}) \). The mutually orthogonal property allows us to decompose \( \tilde{X} \) column-wisely to \( m \) subsets \( \tilde{X}^{(i)}, i = 1, 2, \ldots, m \), and still retain consistency if one fits a linear regression on each subset. To see this, notice that each sub-model now takes a form of \( \tilde{Y} = \tilde{X}^{(i)} \beta^{(i)} + \tilde{W}^{(i)} \) where \( \tilde{W}^{(i)} = \tilde{X}^{(-i)} \beta^{(-i)} + \tilde{\varepsilon} \) and \( \tilde{X}^{(-i)} \) stands for variables not included in the \( i^{th} \) subset. If, for example, we would like to compute the ordinary least squares estimates, it follows

\[
\hat{\beta}^{(i)} = (\tilde{X}^{(i)T} \tilde{X}^{(i)})^{-1} \tilde{X}^{(i)} \tilde{Y} = \beta^{(i)} + (\tilde{X}^{(i)T} \tilde{X}^{(i)})^{-1} \tilde{X}^{(i)} \tilde{W}^{(i)} = \beta^{(i)} + (\tilde{X}^{(i)T} \tilde{X}^{(i)})^{-1} \tilde{X}^{(i)} \tilde{\varepsilon} = \beta^{(i)} + O_p \left( \frac{1}{\sqrt{n}} \right)
\]

where we retrieve a consistent estimator that converges in a rate as if the full dataset were used.

When \( p \) is larger than \( n \), the decorrelation matrix becomes identical to the preconditioning matrix \( F = (XX^T)^{-\frac{1}{2}} \) introduced in Section 2.4.1, i.e., the decorrelated data takes a new form as

\[
\frac{(XX^T/p)^{\frac{1}{2}} Y}{\sqrt{p} U V^T \beta + (XX^T/p)^{\frac{1}{2}} \varepsilon}.
\]
The key difference compared to the low dimensional example is that the new features \( \tilde{X} \) are no longer exactly orthogonal to each other, because of the high dimensionality. Nevertheless, as shown in Theorem 2.4, the correlations between different columns are still roughly of the order \( \sqrt{\frac{\log p}{n}} \) for elliptical families, making the new features approximately orthogonal when \( \log(p) = o(n) \). This allows us to follow the same strategy of partitioning the feature space as in the low dimensional case.

The concrete DECO framework consists of two main steps. Assume \( X \) has been partitioned column-wisely into \( m \) subsets \( X^{(i)}, i = 1, 2, \ldots, m \) and distributed onto \( m \) machines with \( Y \). In the first stage, we obtain the decorrelation matrix \( (XX^T/p)^{-1/2} \) or \( (XX^T/p + rI_n)^{-1/2} \) (for robustness) by computing \( XX^T \) in a distributed way as \( XX^T = \sum_{i=1}^{m} X^{(i)}X^{(i)T} \) and perform the SVD decomposition on \( XX^T \) on a central machine. In the second stage, each worker receives the decorrelation matrix, multiplies it to the local data \( (Y, X^{(i)}) \) to obtain \( (\tilde{Y}, \tilde{X}^{(i)}) \), and fits a penalized regression. When the model is assumed to be exactly sparse, we can potentially apply a refinement step by re-estimating coefficients on all the selected variables simultaneously on the master machine via ridge regression. The details are provided in Algorithm 4.

Line 12 - 14 in Algorithm 4 are added only for the data analysis in Section 5.3, in which \( p \) is so massive that \( \log(p) \) would be comparable to \( n \). For such extreme cases, the algorithm may not scale down the size of \( p \) sufficiently for even obtaining a ridge regression estimator afterwards. Thus, a further sparsification step is recommended. This differs fundamentally from the merging step in SAM (Song and Liang, 2014) in that DECO needs this step only for extreme cases where \( \log(p) \sim n \), while SAM always requires a merging step regardless of the relationship between \( n \) and \( p \). The condition in Line 16 is barely triggered in our experiments (only in Section 5.3), but is crucial for improving the performance for extreme cases. In Line 6, the algorithm
Algorithm 4 The DECO framework

**Initialization:**
1: Input \((Y, X), p, n, m\). Standardize \(X\) and \(Y\) to \(x\) and \(y\) with mean zero;
2: Partition (arbitrarily) \((y, x)\) into \(m\) disjoint subsets \((y, x^{(i)})\) and distribute to \(m\) machines;

**Stage 1: Decorrelation**
3: \(\bar{F} = 0\) initialized on the master machine;
4: for \(i = 1\) to \(m\) do
5: \(F = F + x^{(i)}x^{(i)T}\);
6: \(\bar{F} = \sqrt{p}(F + r_1I_p)^{-1/2}\) on the master machine and then pass back;
7: for \(i = 1\) to \(m\) do
8: \(\tilde{y} = \bar{F}y\) and \(\tilde{x}^{(i)} = \bar{F}x^{(i)}\);

**Stage 2: Estimation/feature selection**
9: for \(i = 1\) to \(m\) do
10: Select features \(\hat{\gamma}^{(i)}\) using either Algorithm 1 or 2;
11: \(\hat{\gamma} = (\hat{\gamma}^{(1)}, \hat{\gamma}^{(2)}, \ldots, \hat{\gamma}^{(m)})\) on the master machine;

**Stage 3: Refinement (optional)**
12: if \(|\hat{\gamma}| \geq n\) then
13: # Sparsification is needed before ridge regression.
14: Select \(\tilde{\gamma}\) from \(\hat{\gamma}\) using either Algorithm 1 or 2;
15: Estimate \(\hat{\beta}_s\) using ridge regression or Bayesian inference;
16: return \(\hat{\beta}\);

inverts \(XX^T + r_1I\) instead of \(XX^T\) for robustness, because the rank of \(XX^T\) after standardization will be \(n-1\). Using ridge refinement instead of ordinary least squares is also for robustness. The precise choice of \(r_1\) and \(r_2\) will be discussed in the numerical section.

Penalized regression fitted using regularization path usually involves a computational complexity of \(O(knp + kd^2)\), where \(k\) is the number of path segmentations and \(d\) is the number of features selected. Although the segmentation number \(k\) could be as bad as \((3^p + 1)/2\) in the worst case (Mairal and Yu, 2012), real data experience suggests that \(k\) is on average \(O(n)\) (Rosset and Zhu, 2007), thus the complexity for DECO takes a form of \(O(n^3 + n^2 \frac{p}{m} + m)\) in contrast to the full lasso which takes a form of \(O(n^2p)\). As \(n\) is assumed to be small, using DECO can substantially reduce the computational cost if \(m\) is properly chosen.
4.3 Theoretical properties

In this section, we provide theoretical justification for the DECO framework using the regularized regression or Bayesian variable selection method. In particular, we prove the consistency for the estimator obtained after Stage 2 of DECO, while the consistency of Stage 3 will then follow immediately. Recall that DECO works with the decorrelated data $\tilde{X}$ and $\tilde{Y}$, which are distributed on $m$ different machines, i.e.,

$$\tilde{Y} = \tilde{X}^{(i)} \beta^{(i)} + \tilde{W}^{(i)}, \quad \text{and} \quad \tilde{W}^{(i)} = \tilde{X}^{(-i)} \beta^{(-i)} + \varepsilon,$$

where $X^{(-i)}$ stands for variables not included in the $i^{th}$ subset. Therefore, it suffices for us to verify the conditions in Theorem 2.2 and 2.3 for all pairs $(\tilde{Y}, \tilde{X}^{(i)})$, $i = 1, 2, \cdots, m$. As previously discussed, partitioning the feature space will result in mis-specified models on each subset, as some strong signals are allocated to the the other subsets. Thus, the key to a consistent feature selection relies on capturing the behavior of the "noise" $\|\tilde{X}^{(i)^T} \tilde{W}^{(i)}\|_x = \|\tilde{X}^{(-i)^T} \tilde{X}^{(-i)} \beta^{(-i)} + \tilde{X}^{(-i)^T} \varepsilon\|_x$, for which we have the following result.

**Lemma 3.** Assume $X \sim EN(\Sigma)$. If $p > c_0 n$ for some $c_0 > 1$, then have for any $\gamma_{11}, \gamma_{12} \subseteq [p] \text{ with } \gamma_{11} \cap \gamma_{12} = \emptyset \text{ and any } t > 0$, we have

$$\mathbb{P}\left( \frac{1}{n} \|\tilde{X}^T \tilde{X}_{\gamma_{12}} \beta_{\gamma_1} \|_x \leq \frac{c_4 e^s \|\beta_{\gamma_1}\|_2 t}{c_s \sqrt{n}} \right) \geq 1 - 2|\gamma_1|e^{-t^2/2} - 5|\gamma_1|e^{-C_n},$$

The proof is provided in Appendix A.3. Lemma 3 along with Theorem 2.5 and 2.6 provides a bound on the biased noise from the mis-specified models. Denote the feature set on $i^{th}$ subset by $\gamma^{(i)}$. We are now ready to state the main results.

**4.3.1 The convergence rate for regularized regression**

Similar as Section 3.2, we will prove the consistency result for individual subsets. By a straightforward concatenating, the consistency of the combined models will follow immediately. For regularized regression, we have the following result.
Theorem 4.1 (Convergence with Gaussian noise). Assume $X \sim N(0, \Sigma)$, $\varepsilon \sim N(0, \sigma^2)$. We also assume $p > c_0 n$, $b_* \leq B \sigma \sqrt{\log p \over n}$ and $\|\beta_*\|_2 \leq M_0 \sigma$, where $M_0^2 \leq \frac{d^2 n}{\sigma^2 \log p}$. Now choosing $\lambda = c \sigma \sqrt{\log p \over n}$ and there exist some absolute constants $V_1, V_2, V_3$ and $C_3$ such that if the following hold,

$$c \geq 12 V_1 M_0 + 6 C_3 + 2 V_2 B, \quad d_* \geq \left( 8 C_3 V_3 + 2 V_1 V_3 M_0 + 4 e V_3 \right) \sigma \sqrt{\log p \over n}$$

and

$$n \geq \max \left\{ \frac{16}{9} c^2 V_3^2, \frac{12^2 V_1^2 V_3^2 s_0^2}{\sigma^2}, \frac{4 \kappa_0 V_3^2}{\sigma} \right\} \log p,$$

then with probability at least $1 - 17 p^{-1}$, the solution of Algorithm 1 satisfies that for each $i \in [m]$

$$\|\hat{\beta}^{(i)}_{\gamma_*} - \beta^{(i)}_{\gamma_*} \|_\infty \leq \left( 4 C_3 V_3 \sigma + V_1 V_3 M_0 + 2 V_3 c \sigma \right) \sqrt{\log p \over n} \quad \text{and} \quad \hat{\beta}_{\gamma_*}^{(i)} = 0$$

and thus the concatenated solution $\hat{\beta}$ satisfies

$$\|\hat{\beta}_{\gamma_*} - \beta_{\gamma_*} \|_\infty \leq \left( 4 C_3 V_3 \sigma + V_1 V_3 M_0 + 2 V_3 c \sigma \right) \sqrt{\log p \over n} \quad \text{and} \quad \hat{\beta}_{\gamma_*} = 0,$$

The constants are defined as $V_1 = \frac{c_4 c_*}{c_*}$, $V_2 = \frac{2 c_*}{c_*}$, $V_3 = \frac{c_*}{c_1 c_*}$ and $c_1, c_2, c_4, C_3$ are constants defined in Theorem 2.4 and 2.6 and $c_*, c^*$ are the smallest and the largest eigenvalues of $\Sigma$.

Proof. The verification of (1) and (2) will be the same as in Theorem 3.1, i.e., with probability at least $1 - 5 p^2 e^{-C n} - 2 p^{-1}$ we have $\tau_1 = \frac{3 c_4 c_*}{\sigma c_*}$ and with probability at least $1 - 4 e^{-C n}$ we have $M_1 = \frac{c_1 c_*}{c_*}$ and $M_2 = \frac{2 c_*}{c_*}$ for all subsets. Now for $\|X^{(i)^T} \hat{W}^{(i)}\|$ we have

$$\|X^{(i)^T} \hat{W}^{(i)}\|_\infty = \|X^{(i)^T} X^{(i)} \beta^{(-)} + X^{(i)^T} \tilde{\varepsilon}\|_\infty.$$
Using Theorem 2.6 we have $\frac{1}{n}\|\tilde{X}(-i)T\tilde{e}\|_{\infty} \leq 2C_3\frac{\sigma_2 \log p}{\sqrt{n}}$ with probability at least $1 - 2pe^{-Cn} - 4p^{-1}$. Taking $t = 2\sqrt{\log p}$ in Lemma 3 with $\gamma_1 = \gamma^{(i)}$ and $\gamma_2 = (\gamma^{(i)})^c$, we have $\frac{1}{n}\|\tilde{X}^{(i)}_i \tilde{X}(-i)^T_\gamma \|_{\infty} \leq \frac{2c_4 e^s \beta^{(i)}_{*}}{c_*} \sqrt{\frac{\log p}{n}}$ with probability at least $1 - 2|\gamma^{(i)}|p^{-2} - 5|\gamma^{(i)}|e^{-Cn}$. Thus we have

$$\max_i \frac{1}{n}\|\tilde{X}^{(i)}_i \tilde{W}_i\|_{\infty} \leq 2C_3\sigma \sqrt{\frac{\log p}{n}} + \frac{2c_4 e^s \beta^{(i)}_{*}}{c_*} \sqrt{\frac{\log p}{n}}$$

with probability at least $1 - pe^{-Cn} - 4p^{-1} - (\sum_{i=1}^{m} 2|\gamma^{(i)}|p^{-2} + 5|\gamma^{(i)}|e^{-Cn}) = 1 - 6p^{-1} - 6pe^{-Cn}$. This indicates that $\tau_2 = 2C_3 + \frac{2c_4 e^s M_0}{c_*}$. To use Theorem 2.2, we also need to bound $\tilde{X}_\gamma^{(i)} \tilde{X}_\gamma^{(i)} c_{\gamma^{(i)}(\gamma^{(i)})}$, here $X^{(i)}_\gamma$ stands for the strong features restricted to the $i$th subset and similar for $X^{(i)}_\gamma$. Using Lemma 3 again with $\gamma_1 = \gamma^{(i)}$ and $\gamma_2 = (\gamma^{(i)})^c$ and taking $t = 2\sqrt{\log p}$, we have

$$\mathbb{P}\left(\frac{1}{n}\|\tilde{X}^{(i)}_\gamma \tilde{X}^{(i)}_\gamma c_{\gamma^{(i)}(\gamma^{(i)})}\|_{\infty} \leq \frac{2c_4 e^s \beta^{(i)}_{\gamma^{(i)}}}{c_*} \sqrt{\frac{\log p}{n}}\right) \geq 1 - 2|\gamma^{(i)}|p^{-2} - 5|\gamma^{(i)}|e^{-Cn},$$

and thus we have for probability greater than $1 - 2|\gamma^{(i)}|p^{-2} - 5|\gamma^{(i)}|e^{-Cn}$

$$\max_{i \in \{1, 2, \cdots, m\}} \frac{1}{n}\|\tilde{X}^{(i)}_\gamma \tilde{X}^{(i)}_\gamma c_{\gamma^{(i)}(\gamma^{(i)})}\|_{\infty} \leq \frac{2c_4 e^s M_0}{c_*} \sqrt{\frac{\log p}{n}},$$

showing that $\tau_3 = \frac{2c_4 e^s M_0}{c_*}$. Now, applying Theorem 2.2 we get if $c \geq \frac{12c_4 e^s M_0}{c_*} + 6C_3 + \frac{2c_4 e^s B}{c_*}$ and $n \geq \max\left\{\frac{16c^2 e^s}{9c^2 c_*}, \frac{12^2 c^2 e^s}{\sigma^2 c^2 c_*}, \frac{4 c^2 e^s}{c^2 c_*}\right\} \log p$ and $d_* \geq \left(\frac{8C_3 c^*}{c_1 c_*} + \frac{2c_4 e^s M_0}{c_1 c_*}\right) + \frac{4 c^*}{c_1 c_*} \sqrt{\frac{\log p}{n}}$, then for $p > 7$, with probability at least $1 - 7p^2 e^{-Cn} - 10p^{-1}$

$$\max_{i \in \{1, 2, \cdots, m\}} \|\tilde{\beta}^{(i)}_{\gamma^{(i)}} - \beta^{(i)}_{\gamma^{(i)}}\|_{\infty} \leq \left(\frac{8C_3 c^*}{2c_1 c_*} + \frac{2c_4 e^s M_0}{c_1 c_*} + \frac{4 c^*}{2c_1 c_*}\right) \sqrt{\frac{\log p}{n}} \quad \text{and} \quad \tilde{\beta}^{(i)}_{\gamma^{(i)}} = 0, \forall i \in \{1, 2, \cdots, m\}.$$

Immediately, we obtain the result on the concatenated coefficient $\tilde{\beta}$ as

$$\|\tilde{\beta}^{(i)}_{\gamma^{(i)}} - \beta^{(i)}_{\gamma^{(i)}}\|_{\infty} \leq \left(\frac{8C_3 c^*}{2c_1 c_*} + \frac{2c_4 e^s M_0}{c_1 c_*} + \frac{4 c^*}{2c_1 c_*}\right) \sqrt{\frac{\log p}{n}} \quad \text{and} \quad \tilde{\beta}^{(i)}_{\gamma^{(i)}} = 0.$$
and for the heavy-tailed noise we have the following result.

**Theorem 4.2** (Convergence with heavy-tailed noise). Assume $X \sim EN(\Sigma)$, $\text{var}(\varepsilon) = \sigma^2$, $\mathbb{E}[L^{-2}] = l_2$. We also assume $p > c_0n$, $b_* \leq B\sigma\sqrt{\frac{\log p}{n}}$ and $\|\beta_*\|_2 \leq M_0\sigma$, where $M_0^2 \leq \frac{d_2^2 n}{2\sigma^2 \log p}$. Now choosing $\lambda = c\sigma\sqrt{\frac{\log p}{n}}$ and there exist some absolute constants $V_1, V_2, V_3$ and $C_3$ such that if the following hold,

$$c \geq 12V_1M_0 + 3\sqrt{C_1n^\delta} + 2V_2B, \quad d_* \geq \left(4\sqrt{C_1n^\delta}V_3 + 2V_1V_3M_0 + 4cV_3\right)\sigma\sqrt{\frac{\log p}{n}}$$

and

$$n \geq \max\left\{\frac{16}{9}c^2V_3^2, \frac{12^2 V_1^2 V_3^2}{\sigma^2}, \frac{4\kappa_0 V_3^2}{\sigma^2}\right\} \log p,$$

where $\delta \in (0, 1)$ is an arbitrary constant, then with probability at least $1 - 6\log p/n^\delta - 17p^{-1}$, the solution of Algorithm 1 satisfies that for each $i \in [m]$

$$\|\hat{\beta}^{(i)}_{\gamma_*} - \beta^{(i)}_{\gamma_*}\|_\infty \leq 2\sqrt{C_1V_3}\sigma\sqrt{\frac{\log p}{n^{1-\delta}}} + \left(V_1V_3M_0 + 2V_3c\sigma\right)\sqrt{\frac{\log p}{n}} \quad \text{and} \quad \hat{\beta}^{(i)}_{(\gamma_*)^c} = 0$$

and thus the concatenated solution $\hat{\beta}$ satisfies

$$\|\hat{\beta}_{\gamma_*} - \beta_{\gamma_*}\|_\infty \leq 2\sqrt{C_1V_3}\sigma\sqrt{\frac{\log p}{n^{1-\delta}}} + \left(V_1V_3M_0 + 2V_3c\sigma\right)\sqrt{\frac{\log p}{n}} \quad \text{and} \quad \hat{\beta}_{(\gamma_*)^c} = 0,$$

The constants are defined as $V_1 = \frac{c_1c_*}{c_n^*}, V_2 = \frac{c_2c_*}{c_n^*}, V_3 = \frac{c}{c_1c_*}$ and $c_1, c_2, c_4, C_3$ are constants defined in Theorem 2.4 and 2.6 and $c_*, c^*$ are the smallest and the largest eigenvalues of $\Sigma$.

**Proof.** With heavy-tailed noise we need to update the value of $\tau_2$ for the $\|\tilde{X}^{(i)}T\tilde{\varepsilon}\|_\infty$ part. According to the calculation in Theorem 3.2, we just need to replace $2C_3$ by
\( \sqrt{C_1 n^8} \), i.e., with probability at least \( 1 - 6 \log p/n^6 - 6p^{-1} - 7pe^{-Cn} \)

\[ \tau_2 = \sqrt{C_1 n^8} + \frac{2c_4c^*M_0}{c_*} \]

holds for all subsets. The rest will follow exactly the same as the proof for Theorem 4.1. with probability at least \( 1 - 5p^2e^{-Cn} - 2p^{-1} \) we have \( \tau_1 = \frac{3c_4c^*}{\sigma c_*} \) and with probability at least \( 1 - 4e^{-Cn} \) we have \( M_1 = \frac{c_4c^*}{c_*} \) and \( M_2 = \frac{c_4c^*}{c_*} \) for all subsets. In addition, for probability greater than \( 1 - 2|\gamma_*|p^{-2} - 5|\gamma_*|e^{-Cn} \), we have \( \tau_3 = \frac{2c_4c^*M_0}{c_*} \). Therefore, we have if \( c \geq \frac{12c_4c^*M_0}{c_*} + 3\sqrt{C_1 n^8} + 2c_4c^*B/c_* \) and \( n \geq \max \left\{ \frac{16c^2c^*}{c^2c_*}, \frac{12^2c^2c^*\delta_0^2}{\sigma^2c^2c_*}, \frac{4c^2c^*}{c^2c_*} \right\} \) \log p

and \( d_* \geq \left( \frac{4\sqrt{C_1 c^*}}{c_1 c_*} n^\delta + \frac{2c_4c^*^2M_0}{c_1 c_*^2} + \frac{4cc^*}{c_1 c_*} \right) \sigma \sqrt{\frac{\log p}{n}} \), then for \( p > 7 \), with probability at least \( 1 - 6 \log p/n^6 - 17p^{-1} \)

\[
\max_{i \in [m]} \| \hat{\beta}^{(i)}_{\gamma_*} - \beta^{(i)}_{\gamma_*} \|_2 \leq \left( \frac{4\sqrt{C_1 c^*}}{2c_1 c_*} n^\delta + \frac{2c_4c^*^2M_0}{c_1 c_*^2} + \frac{4cc^*}{2c_1 c_*} \right) \sqrt{\frac{\log p}{n}} \quad \text{and} \quad \hat{\beta}^{(i)}_{(\gamma_*)c} = 0, \forall i \in [m].
\]

Immediately, we obtain the result on the concatenated coefficient \( \hat{\beta} \) as

\[
\| \hat{\beta}_{\gamma_*} - \beta_{\gamma_*} \|_2 \leq \left( \frac{4\sqrt{C_1 c^*}}{2c_1 c_*} n^\delta + \frac{2c_4c^*^2M_0}{c_1 c_*^2} + \frac{4cc^*}{2c_1 c_*} \right) \sqrt{\frac{\log p}{n}} \quad \text{and} \quad \hat{\beta}_{(\gamma_*)c} = 0.
\]

The two theorems show that by using DECO framework, not only the final output will be consistent but also the feature selection within each subset will also be consistent regardless of the mis-specification. We have similar results for Bayesian variable selection in next section.

4.3.2 The convergence rate for Bayesian variable selection

For Bayesian variable selection, the result is quite similar to the regularized regression. We note that although partitioned subsets have smaller number of features, we
still apply the same prior in Condition 3 to all subsets with \( p \) being the total number of features. We have the following theorem for Gaussian noise and heavy-tailed noise.

**Theorem 4.3** (Convergence with Gaussian noise). Assume \( X \sim N(0, \Sigma), \varepsilon \sim N(0, \sigma^2) \). If \( p > c_0 n, \|\beta\|_1 \leq M_0, b_* \leq B \sigma \sqrt{\frac{\log p}{n}} \) and \( \|\beta_*(\gamma_*)^\nu\|_1 \leq R \), then there exist some constants \( V'_0, V'_1, V'_2, V'_3, V'_4 \) and \( V'_5 \) such that if the following conditions hold

\[
\alpha \geq V'_0 + V'_1 M^2_0 + 18V'_2 V'_4 \sigma^{-2} R^2 + 3 - \kappa,
\]

\[
\frac{d^*}{\alpha + \kappa + 5} \geq V'_3 M^2_0 \left( \frac{4BR \sqrt{\log p}}{M^2_0} \sqrt{\frac{\log p}{n}} + V'_5 + \frac{7}{p^2} \right) \frac{\sigma^2 \log p}{n}
\]

and

\[
n \geq \max \left\{ \frac{1}{2} V'_0 s'_0 + \frac{1}{2} V'_1 s'_0 M^2_0, V'_2 s'_0^2, 72\sigma^{-2}(V'_0 + V'_1 M^2_0) V'_4 R^2 \right\} \log p
\]

then with probability at least \( 1 - 16p^{-1} - 2e^{-n/4} \),

\[
\min_{i \in [m]} \pi_n(\gamma_*^{(i)}|Y) \geq 1 - 3p^{-2},
\]

for \( p \geq \max\{\frac{12B R \nu}{M^2_0} \sqrt{\frac{\log p}{n}} + \frac{33\nu c^2}{c_1 c^2 v_1 M^2_0} + 12c_1 c_2 c^3 \nu^{1/\alpha}, 9\} \). The constants are defined as

\[
V'_0 = \frac{29c^2 c^2 v_2 v_2}{c_1 c^*}, V'_1 = \frac{29c^2 c^2 v_2 v_2}{c_1 c^*}, V'_2 = \frac{36c^2 c^2 v_2}{c_1 c^*}, V'_3 = \frac{64c^4 v_2}{c^2 c^*}, V'_4 = c_1 c^* v_2 \text{ and}
\]

\[
V'_5 = \frac{11}{c_1 c^2 v_1 M^2_0} + \frac{c^2}{c_1 c^*}, \text{ where } c_1, c_2, c_4, c^*, v_1, C_3 \text{ and } v_2 \text{ are defined in Theorem 2.4, 2.6 and Proposition 4.}
\]

**Proof.** In addition to all the conditions verified in Theorem 4.1 and 3.4, we also have to quantify \( \frac{1}{n} \|\tilde{W}^{(i)}\|_2^2 \) which can be expressed as

\[
\frac{1}{n} \|\tilde{W}^{(i)}\|_2^2 = \frac{1}{n} \|\tilde{\varepsilon}\|_2^2 + \frac{1}{n} \|\tilde{X}^{(-i)} \beta_*(\gamma_*)^{\nu} - \beta_*(\gamma_*)^{\nu}\|_2^2 + \frac{2}{n} \varepsilon^T \tilde{X}^{(-i)} \beta_*(\gamma_*)^{\nu},
\]

where the first term can be bounded by using Theorem 2.6, the second term can be bounded by using Theorem 2.4 and the last term follows

\[
\frac{2}{n} |\varepsilon^T \tilde{X}^{(-i)} \beta_*(\gamma_*)^{\nu} - \beta_*(\gamma_*)^{\nu}| \leq \frac{2}{n} \|\varepsilon^T \tilde{X}^{(-i)}\|_2 \|\beta_*(\gamma_*)^{\nu} - \beta_*(\gamma_*)^{\nu}\|_1
\]

53
Therefore, we have uniformly for all $i \in [m]$

$$\frac{1}{n} \| \tilde{W}^{(i)} \|_2^2 \geq \frac{\sigma^2}{2c^* \nu_2} - 4C_3\sigma^2 M_0 \sqrt{\log p \over n}$$

and using the fact $\| \beta^{(i)}_* \|_2 \leq \| \beta^{(i)}_* \|_1$ we have

$$\frac{1}{n} \| \tilde{W}^{(i)} \|_2^2 \leq \frac{5\sigma^2}{2c^* \nu_1} + \frac{c_2c^*}{c_*} M_0^2 \sigma^2 + 4C_3\sigma^2 M_0 \sqrt{\log p \over n}$$

with probability $1 - 3pe^{-CN} - 5p^{-1} - 2e^{-n/4}$. This means that $r' = \frac{1}{4c^* \nu_2}$ and $r = \frac{11}{4c^* \nu_1} + \frac{c_2c^*}{c_*} M_0^2$ as long as $M_0^2 \leq \frac{n}{64c_2c^*^2C_3^2\log p}$. The above event along with $\tau_2 = 2C_3 + \frac{2c_2c^* M_0}{c_*}$ has probability greater than $1 - 7pe^{-CN} - 7p^{-1} - 2e^{-n/4}$. $q_1$ and $q_2$ can choose the same values as in Theorem 3.4,

$$q_2 = \frac{c_1c^*^2 M_0^2}{c_* p^2} \quad \text{and} \quad q_1 = \frac{c_1c^*^2 BR\sigma}{c_*} \sqrt{\log p \over n}$$

with probability at least $1 - 4e^{-CN}$. In addition, we have $\tau_1 = \frac{3c_4c^*/\sigma c_*}$ with probability at least $1 - 5p^2e^{-CN} - 2p^{-1}$ and $M_1 = \frac{c_1c^*}{c_*}$, $M_2 = \frac{2c^*c}{c_*}$ with probability at $1 - 4pe^{-CN}$.

By summarizing all results and taking $C = 3$ we obtain that if

$$\alpha \geq \frac{2^9c_4^2C_2^2\nu_2}{c_1c_*} + \frac{2^9c_4^2c^*^4 \nu_2 M_0^2}{c_1c_*^3} + \frac{468c_4^2c^*^4 \nu_2 R^2}{c_1c_*^3 \sigma^2} + 2 - \kappa$$

and

$$A^2 \over \alpha + \kappa + 5 \geq \frac{64c^*^4 \nu_2 M_0^2}{c_*^2} \left( \frac{4BR \sqrt{\log p \over n}}{M_0^2} + \frac{7p^{-\alpha}}{c_1c^*^2 \nu_1 M_0^2} + 11 \over c_1c^*^2 \nu_1 M_0^2 + 1 \over c_1c_*^* \right)$$

and

$$n \geq \max \left\{ \frac{2^8c_4^2c^*_4 \nu_2 \rho_0}{c_1c_*} + \frac{2^8c_4^2c^*^4 \nu_2 M_0^2 \rho_0}{c_1c_*^3} \over 36c^*_4c^*^4 \rho_0^2 \sigma^2 R^2 \over c_*^2} + \frac{9 \times 2^{12}c^*_4c^*^4 \nu_2 M_0^2 \rho_0}{c^*_4 \sigma^2} \right\} \log p$$

54
and
\[ p^\alpha \geq \frac{12BR\sigma}{M_0^2} \sqrt{\frac{\log p}{n}} + \frac{33}{c_1 c^* v_1 M_0^2} + 12 c_1 c_2 e^3, \]
then for \( p \geq 9 \) with probability at least \( 1 - 16p^{-1} - 2e^{-n/4} \),
\[ \min_{i \in [m]} \pi_n(\gamma_{(i)}^*|Y) \geq 1 - 3p^{-2}. \]

We also have the result for heavy-tailed noise.

**Theorem 4.4** (Convergence with heavy-tailed noise). Assume \( x^{(k)} \sim EN(L_k, \Sigma) \), \( \text{var}(\varepsilon_k) = \sigma^2 \), \( \mathbb{E}[L_k^2] = l_2 \) and \( \mathbb{E}[\varepsilon_k^4/L_k^2] \leq M_4, k \in [n] \). If \( p > c_0 n \), \( \|\beta_*\|_1 \leq M_0 \sigma \), \( b_* \leq B\sigma \sqrt{\frac{\log p}{n}} \) and \( \|\beta_*(\gamma_*)\|_1 \leq R \), then there exist some constants \( V_0'', V_1'', V_2', V_3', V_4'' \) and \( V_5'' \) such that if the following conditions hold

\[ \alpha \geq V_0'' n^\delta + V_1'' M_0^2 + 27 V_2'' V_4'' \sigma^{-2} R^2 + 3 - \kappa, \]
\[ \frac{d_*^2}{\alpha + \kappa + 5} \geq V_3'M_0^2 \left( \frac{4BR\sigma}{M_0^2} \sqrt{\frac{\log p}{n}} + V_5'' + \frac{7}{p^\alpha} \right) \frac{\sigma^2 \log p}{n} \]

and

\[ n \geq \max \left\{ \frac{1}{2} V_0'' s_0 n^\delta + \frac{1}{2} V_1'' s_0 M_0^2, \ V_2's_0^2, \ 108 \sigma^{-2} (V_0'' n^\delta + V_1'' M_0^2) V_4'' R^2 \right\} \log p \]

where \( \delta \in (0,1) \) is an arbitrary constant, then with probability at least \( 1 - 6 \log p/n^\delta - M_4/n - 17p^{-1} \),
\[ \min_{i \in [m]} \pi_n(\gamma_{(i)}^*|Y) \geq 1 - 3p^{-2}, \]
for \( p \geq \max \left\{ \left( \frac{12BR\sigma}{M_0^2} \sqrt{\frac{\log p}{n}} + \frac{60}{c_1 c^* v_1 M_0^2} + 12 c_1 c_2 e^3 \right)^{1/\alpha}, 9 \right\} \). The constants are defined as \( V_0'' = \frac{192c^2 s^2 C^2 v_2}{c_1 c_*}, V_1'' = \frac{3 \times 2^6 c^2 s^4 v_2}{c_1 c_*^3}, V_2' = \frac{36 c^2 c^*}{c_1 c_*}, V_3' = \frac{64 c^4 v_2}{c_*}, V_4'' = \frac{c_1 c_*}{c^* v_1 l_1}, \) and
\[ V_n^* = \frac{20v_1l_1}{c_1 c^2 v_1 M_0} + \frac{c_2}{c_1 c^*} \text{ where } c_1, c_2, c_4, c^*, v_1, C_3 \text{ and } v_2 \text{ are defined in Theorem } 2.4, 2.6 \text{ and Proposition } 4. \]

**Proof.** The proof is essentially the same as Theorem 4.3, with updated values for \( r' \), \( r \) and \( \tau_2 \). For \( \tau_2 \), we have with probability at least \( 1 - 6 \log p/n^\delta - 7p^{-1} - 7p e^{-Cn} \)

\[ \tau_2 = \sqrt{C_1 n^\delta + \frac{2c_4 c^* M_0}{c^*}} \]

holds for all subsets. According to the Theorem 3.5 we know \( \frac{1}{4} v_1 l_1 \leq \frac{1}{n} \| \hat{W} \|_2^2 \leq 4v_2 l_2 \) with probability \( 1 - M_4/n - 4p^{-2} \). Thus for \( \frac{1}{n} \| \hat{W}^{(i)} \|_2^2 \) we have uniformly for all \( i \in [m] \)

\[ \frac{1}{n} \| \hat{W}^{(i)} \|_2^2 \geq \frac{v_1 l_1}{4} \sigma^2 - 2\sqrt{C_1 \sigma^2 M_0} \sqrt{\frac{\log p}{n^{1-\delta}}} \]

and

\[ \frac{1}{n} \| \hat{W}^{(i)} \|_2^2 \leq 4v_2 l_2 \sigma^2 + \frac{c_2 c^*}{c^*} M_0^2 \sigma^2 + 2\sqrt{C_1 \sigma^2 M_0} \sqrt{\frac{\log p}{n^{1-\delta}}} \]

which is true along with the value of \( \tau_2 \) for probability at least \( 1 - 6 \log p/n^\delta - M_4/n - 8p^{-1} - 7p e^{-Cn} \), i.e., \( r' = \frac{v_1 l_1}{6} \) and \( r = 5v_2 l_2 + \frac{c_2 c^*}{c^*} M_0^2 \) as long as \( M_0^2 \leq \frac{v_2^2 \sigma^2 n^{1-\delta}}{576c_1 \log p} \). The values for \( q_1 \) and \( q_2 \) are the same as in Theorem 4.3, i.e.,

\[ q_2 = \frac{c_1 c^2 M_0^2}{c^* p^2} \quad \text{and} \quad q_1 = \frac{c_1 c^2 B R \sigma}{c^*} \sqrt{\frac{\log p}{n}} \]

with probability at least \( 1 - 4e^{-Cn} \). In addition, we have \( \tau_1 = \frac{3c_4 c^*}{c c^*} \) with probability at least \( 1 - 5p^2 e^{-Cn} - 2p^{-1} \) and \( M_1 = \frac{c_4 c^*}{c^*} \), \( M_2 = \frac{c_2 c^*}{c^*} \) with probability at least \( 1 - 4p e^{-Cn} \).

With these new values, we can state now our result. Taking \( C = 3 \) we obtain that if

\[ \alpha \geq \frac{3 \times 2^6 c^* C_1 n^\delta}{c_1 c^* v_1 l_1} + \frac{3 \times 2^8 c_1^2 c^3 M_0^2}{c_1 c^* v_1 l_1} + \frac{3 \times 234 c_1^3 c^3 R^2}{c_1 c^* v_1 l_1 \sigma^2} + 2 - \kappa \]

and

\[ \frac{A^2}{\alpha + \kappa + 5} \geq \frac{64 c^*^4 v_2 M_0^2}{c^*} \left( \frac{4BR}{M_0} \sqrt{\frac{\log p}{n}} + 7p^{-\alpha} + \frac{20v_2 l_2}{c_1 c^2 v_1 M_0^2} + \frac{c_2}{c_1 c^*} \right) \]

56
and

\[
    n \geq \max \left\{ \frac{3 \times 2^2 c^8 C_1 s_0 n^{\delta}}{c_1 c_6 v_1 l_1} + \frac{3 \times 2^7 c_3^2 s_3 M_0^{2 s_0}}{c_1 c_6^3 v_1 l_1}, \frac{3^4 \times 2^3 \sigma_1 R^2 n^{\delta}}{\sigma_1^2 v_1 l_1} + \frac{3^4 \times 2^{10} c_4^2 s_2 M_0^2 R^2}{c_6^2 \sigma^2 v_1^2 l_1}, \log p \right\}
\]

and

\[
    p^a \geq \frac{12 B R \sigma}{M_0^2} \sqrt{\frac{\log p}{n}} + \frac{60 v_2 l_2}{c_1 c^2 s_1 v_1 M_0^2} + 12 c_1 c_2 c^3,
\]

then for \( p \geq 9 \) with probability at least \( 1 - 6 \log p/n^\delta - M_4/n - 17p^{-1} \),

\[
    \min_{i \in [m]} \pi_n (\gamma_1(i) | Y) \geq 1 - 3p^{-2}.
\]

Theorem 4.3 and 4.4 justifies the Bayesian variable selection using \textit{DECO}. In particular, it shows that the posterior probability on all subsets will be concentrated on the strong signal sets. For subsets where no strong signals are allocated, this means that the posterior will concentrate on the null model.

4.4 The relationship with variable screening

In this section, we briefly discuss the relationship between \textit{DECO} and the variable screening. Variable screening is a powerful method first proposed in Fan and Lv (2008) for efficient dimension reduction when the dimensionality is too high. The proposed \textit{SIS} method screens out variables with small marginal correlations and the retained model is then refined via the second-stage feature selection. Thus, the result relies heavily on the property of marginal correlation and could fail when the feature correlations are high. There are a few follow-up research attempting to ameliorate this issue Wang (2009); Wang and Leng (2015) The key for variable screening is that all variables are untangled and treated separately, leading to computational gain
from either the simpler calculation or parallel computation. This is very similar to *DECO*. In the extreme case where we choose \( m = p \), i.e., each individual subset receives only one feature, *DECO* is essentially the same as variable screening.

We specialize our attention to *lasso*. When only one feature is allocated to each subset, *lasso* is solving a simpler optimization problem where the solution can be found in the close form, i.e.,

\[
\hat{\beta}_i = \min_{\beta_i} \frac{1}{n} \| \tilde{Y} - \tilde{x}_i \beta \|_2^2 + \lambda |\beta_i|,
\]

and we have

\[
\hat{\beta}_i = \left( \frac{\tilde{x}_i^T \tilde{Y}}{\|\tilde{x}_i\|_2^2} - \lambda \right)_+,
\]

where \((a)_+ = a \) if \( a \geq 0 \) or \( 0 \) if \( a < 0 \). Thus, *DECO-lasso* shrinks the marginal correlation of the decorrelated data \( \tilde{X} \) and \( \tilde{Y} \) towards zero, which is equivalent to rank the correlation and select the larger ones. We summarized this special procedure in Algorithm 5.

**Algorithm 5** The *DECO-lasso* screening algorithm

<table>
<thead>
<tr>
<th>Initialization:</th>
</tr>
</thead>
<tbody>
<tr>
<td>1: Input ((\tilde{Y}, \tilde{X}), p,n,m). Standardize (X) and (Y) to (x) and (y) with mean zero;</td>
</tr>
<tr>
<td>2: Compute ( F = XX^T; )</td>
</tr>
<tr>
<td>3: ( \tilde{F} = \sqrt{p}(F + r_1 I_p)^{-1/2}; )</td>
</tr>
<tr>
<td>4: ( \tilde{y} = \tilde{F}y ) and ( \tilde{X} = \tilde{F}X; )</td>
</tr>
<tr>
<td>5: for ( i = 1 ) to ( p ) do</td>
</tr>
<tr>
<td>6: Compute ( \hat{\beta}_i = \left( \frac{\tilde{x}_i^T \tilde{Y}}{|\tilde{x}_i|<em>2^2} - \lambda \right)</em>+; )</td>
</tr>
<tr>
<td>7: ( \hat{\gamma} = { k :</td>
</tr>
<tr>
<td>8: return ( \hat{\gamma} );</td>
</tr>
</tbody>
</table>

The *DECO-lasso* is also closely related to another screening algorithm *HOLP* (Wang and Leng, 2015). *HOLP* screens the variables using \( \hat{\beta} = X^T(XX^T)^{-1}Y \) which is equivalent to \( \frac{1}{p} \tilde{X}^T \tilde{Y} \) after decorrelation. Compared to *DECO-lasso*, *HOLP* adopts
the marginal covariance instead of correlation for screening. It remains interesting to see how this difference might lead to different practical performance.

4.5 Experiments

In this section, we use lasso as an example to illustrate the empirical performance of DECO via extensive numerical experiments. In particular, we compare DECO after 2 stage fitting (DECO-2) and DECO after 3 stage fitting (DECO-3) with the full data lasso (lasso-full), the full data lasso with ridge refinement (lasso-refine) and lasso with a naive feature partition without decorrelation (lasso-naive). This section consists of three parts. In the first part, we run DECO-2 on some simulated data and monitor its performance on one randomly chosen subset that contains part of the true signals. In the second part, we verify our claim in Theorem 2.5 and 2.6 that the accuracy of DECO does not depend on the subset number. In the last part, we provide a comprehensive evaluation of DECO’s performance by comparing DECO with other methods under various correlation structures.

The synthetic datasets are from model (4.1) with \( X \sim N(0, \Sigma) \) and \( \varepsilon \sim N(0, \sigma^2) \). The variance \( \sigma^2 \) is chosen such that \( \hat{R}^2 = \text{var}(X\beta)/\text{var}(Y) = 0.9 \). For evaluation purposes, we consider five different structures of \( \Sigma \) as below.

**Model (i) Independent predictors.** The support of \( \beta \) is \( S = \{1, 2, 3, 4, 5\} \). We generate \( X_i \) from a standard multivariate normal distribution with independent components. The coefficients are specified as

\[
\beta_i = \begin{cases} 
(-1)^{Ber(0, 5)} & \left( |N(0, 1)| + 5\sqrt{\frac{\log p}{n}} \right) \\
0 & i \notin S.
\end{cases}
\]

**Model (ii) Compound symmetry.** All predictors are equally correlated with correlation \( \rho = 0.6 \). The coefficients are the same as those in Model (i).
Model (iii) **Group structure.** This example is Example 4 in Zou and Hastie (2005), for which we allocate the 15 true variables into three groups. Specifically, the predictors are generated as \( x_{1+3m} = z_1 + N(0, 0.01) \), \( x_{2+3m} = z_2 + N(0, 0.01) \) and \( x_{3+3m} = z_3 + N(0, 0.01) \), where \( m = 0, 1, 2, 3, 4 \) and \( z_i \sim N(0, 1) \) are independent. The coefficients are set as \( \beta_i = 3, \ i = 1, 2, \ldots, 15; \ \beta_i = 0, \ i = 16, \ldots, p. \)

Model (iv) **Factor models.** This model is considered in Meinshausen and Bühlmann (2010). Let \( \phi_j, j = 1, 2, \ldots, k \) be independent standard normal variables. We set predictors as \( x_i = \sum_{j=1}^k \phi_j f_{ij} + \eta_i \), where \( f_{ij} \) and \( \eta_i \) are independent standard normal random variables. The number of factors is chosen as \( k = 5 \) in the simulation while the coefficients are specified the same as in Model (i).

Model (v) **\( \ell_1 \)-ball.** This model takes the same correlation structure as Model (ii), with the coefficients drawn from Dirichlet distribution \( \beta \sim Dir\left(\frac{1}{p}, \frac{1}{p}, \ldots, \frac{1}{p}\right) \times 10. \) This model is to test the performance under a weakly sparse assumption on \( \beta \), since \( \beta \) is non-sparse satisfying \( \|\beta\|_1 = 10. \)

Throughout this section, the performance of all the methods is evaluated in terms of four metrics: the number of false positives (# FPs), the number of false negatives (# FNs), the mean squared error \( \|\hat{\beta} - \beta_0\|_2^2 \) (MSE) and the computational time (runtime). We use glmnet (Friedman et al., 2010) to fit lasso and choose the tuning parameter using the extended BIC criterion (Chen and Chen, 2008) with \( \gamma \) fixed at 0.5. For DECO, the features are partitioned randomly in Stage 1 and the tuning parameter \( r_1 \) is fixed at 1 for DECO-3. Since DECO-2 does not involve any refinement step, we choose \( r_1 \) to be 10 to aid robustness. The ridge parameter \( r_2 \) is chosen by 5-fold cross-validation for both DECO-3 and lasso-refine. All the algorithms are coded and timed in Matlab on computers with Intel i7-3770k cores. For any embarrassingly parallel algorithm, we report the preprocessing time plus the longest runtime of a single machine as its runtime.
Monitor DECO on one subset

In this part, using data generated from Model (ii), we illustrate the performance of DECO on one randomly chosen subset after partitioning. The particular subset we examine contains two nonzero coefficients $\beta_1$ and $\beta_2$ with 98 coefficients, randomly chosen, being zero. We either fix $p = 10,000$ and change $n$ from 100 to 500, or fix $n$ at 500 and change $p$ from 2,000 to 10,000 to simulate datasets. We fit DECO-2, lasso-full and lasso-naive to 100 simulated datasets, and monitor their performance on that particular subset. The results are shown in Fig 4.1 and 4.2.

It can be seen that, though the sub-model on each subset is mis-specified, DECO
performs as if the full dataset were used as its performance is on par with lasso-full. On the other hand, lasso-naive fails completely. This result clearly highlights the advantage of decorrelation before feature partitioning.

4.5.2 Impact of the subset number $m$

As shown in Theorem 2.5 and 2.6, the performance of DECO does not depend on the number of partitions $m$. We verify this property by using Model (ii) again. This time, we fix $p = 10,000$ and $n = 500$, and vary $m$ from 1 to 200. We compare the performance of DECO-2 and DECO-3 with lasso-full and lasso-refine. The averaged results from 100 simulated datasets are plotted in Fig 4.3. Since $p$ and $n$ are both fixed, lasso-full and lasso-refine are expected to perform stably over different
Figure 4.3: Performance of DECO with different number of subsets.

$m$'s. **DECO-2** and **DECO-3** also maintain a stable performance regardless of the value of $m$. In addition, **DECO-3** achieves a similar performance to and sometimes better accuracy than **lasso-refine**, possibly because the irrepresentable condition is satisfied after decorrelation (See the discussions after Theorem 2.5).

### 4.5.3 Comprehensive comparison

In this section, we compare all the methods under the five different correlation structures. The model dimension and the sample size are fixed at $p = 10,000$ and $n = 500$ respectively and the number of subsets is fixed as $m = 100$. For each model, we simulate 100 synthetic datasets and record the average performance in Table 4.1

Several conclusions can be drawn from Table 4.1. First, when all variables are
Table 4.1: Results for five models with \((n, p) = (500, 10000)\)

<table>
<thead>
<tr>
<th>Model</th>
<th>DECO-3</th>
<th>DECO-2</th>
<th>lasso-refine</th>
<th>lasso-full</th>
<th>lasso-naive</th>
</tr>
</thead>
<tbody>
<tr>
<td>(i)</td>
<td>MSE</td>
<td>0.102</td>
<td>3.502</td>
<td>0.104</td>
<td>0.924</td>
</tr>
<tr>
<td></td>
<td># FPs</td>
<td>0.470</td>
<td>0.570</td>
<td>0.420</td>
<td>0.420</td>
</tr>
<tr>
<td></td>
<td># FNs</td>
<td>0.010</td>
<td>0.020</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>Time</td>
<td>65.5</td>
<td>60.3</td>
<td>804.5</td>
<td>802.5</td>
</tr>
<tr>
<td>(ii)</td>
<td>MSE</td>
<td>0.241</td>
<td>4.636</td>
<td>1.873</td>
<td>3.808</td>
</tr>
<tr>
<td></td>
<td># FPs</td>
<td>0.460</td>
<td>0.550</td>
<td>2.39</td>
<td>2.39</td>
</tr>
<tr>
<td></td>
<td># FNs</td>
<td>0.010</td>
<td>0.030</td>
<td>0.160</td>
<td>0.160</td>
</tr>
<tr>
<td></td>
<td>Time</td>
<td>66.9</td>
<td>61.8</td>
<td>809.2</td>
<td>806.3</td>
</tr>
<tr>
<td>(iii)</td>
<td>MSE</td>
<td>6.620</td>
<td>1220.5</td>
<td>57.74</td>
<td>105.99</td>
</tr>
<tr>
<td></td>
<td># FPs</td>
<td>0.410</td>
<td>0.570</td>
<td>0.110</td>
<td>0.110</td>
</tr>
<tr>
<td></td>
<td># FNs</td>
<td>0.130</td>
<td>0.120</td>
<td>3.93</td>
<td>3.93</td>
</tr>
<tr>
<td></td>
<td>Time</td>
<td>65.5</td>
<td>60.0</td>
<td>835.3</td>
<td>839.3</td>
</tr>
<tr>
<td>(iv)</td>
<td>MSE</td>
<td>0.787</td>
<td>5.648</td>
<td>11.15</td>
<td>6.610</td>
</tr>
<tr>
<td></td>
<td># FPs</td>
<td>0.460</td>
<td>0.410</td>
<td>19.90</td>
<td>19.90</td>
</tr>
<tr>
<td></td>
<td># FNs</td>
<td>0.090</td>
<td>0.100</td>
<td>0.530</td>
<td>0.530</td>
</tr>
<tr>
<td></td>
<td>Time</td>
<td>69.4</td>
<td>64.1</td>
<td>875.1</td>
<td>880.0</td>
</tr>
<tr>
<td>(v)</td>
<td>MSE</td>
<td>—</td>
<td>2.341</td>
<td>—</td>
<td>1.661</td>
</tr>
<tr>
<td></td>
<td>Time</td>
<td>—</td>
<td>57.5</td>
<td>—</td>
<td>829.5</td>
</tr>
</tbody>
</table>

independent as in Model (i), lasso-naive performs similarly to DECO-2 because no decorrelation is needed in this simple case. However, lasso-naive fails completely for the other four models when correlations are presented. Second, DECO-3 achieves the overall best performance. The better estimation error over lasso-refine is due to the better variable selection performance, since the irrepresentable condition is not needed for DECO. Finally, DECO-2 performs similarly to lasso-full and the difference is as expected according to the discussions after Theorem 2.6.

4.5.4 Real data

We illustrate the competitive performance of DECO via three real datasets that cover a range of high dimensionalities, by comparing DECO-3 to lasso-full, lasso-refine and lasso-naive in terms of prediction error and computational time. The algo-
Algorithms are configured in the same way as in Section 4. Although DECO allows arbitrary partitioning (not necessarily random) over the feature space, for simplicity, we confine our attention to random partitioning. In addition, we perform DECO-3 multiple times on the same dataset to ameliorate the uncertainty due to the randomness in partitioning.

**Student performance dataset** We look at one of the two datasets used for evaluating student achievement in two Portuguese schools (Cortez and Silva, 2008). The data attributes include school related features that were collected by using school reports and questionnaires. The particular dataset used here provides the students’ performance in mathematics. The goal of the research is to predict the final grade (range from 0 to 20). The original data set contains 395 students and 32 raw attributes. The raw attributes are recoded as 40 attributes and form 767 features after adding interaction terms. To reduce the conditional number of the feature matrix, we remove features that are constant, giving 741 features. We standardize all features and randomly partition them into 5 subsets for DECO. To compare the performance of all methods, we use 10-fold cross validation and record the prediction error (mean square error, MSE), model size and runtime. The averaged results are summarized in Table 4.2. We also report the performance of the null model which predicts the final grade on the test set using the mean final grade in the training set.

Table 4.2: The results of all methods for student performance data with \((n, p, m) = (395, 741, 5)\)

<table>
<thead>
<tr>
<th>Method</th>
<th>MSE</th>
<th>Model size</th>
<th>Runtime</th>
</tr>
</thead>
<tbody>
<tr>
<td>DECO-3</td>
<td>3.64</td>
<td>1.5</td>
<td>37.0</td>
</tr>
<tr>
<td>lasso-full</td>
<td>3.79</td>
<td>2.2</td>
<td>60.8</td>
</tr>
<tr>
<td>lasso-refine</td>
<td>3.89</td>
<td>2.2</td>
<td>70.9</td>
</tr>
<tr>
<td>lasso-naive</td>
<td>16.5</td>
<td>6.4</td>
<td>44.6</td>
</tr>
<tr>
<td>Null</td>
<td>20.7</td>
<td>—</td>
<td>—</td>
</tr>
</tbody>
</table>
Mammalian eye diseases  This dataset, taken from Scheetz et al. (2006), was collected to study mammalian eye diseases, with gene expression for the eye tissues of 120 twelve-week-old male F2 rats recorded. One gene coded as TRIM32 responsible for causing Bardet-Biedl syndrome is the response of interest. Following the method in Scheetz et al. (2006), 18,976 probes were selected as they exhibited sufficient signal for reliable analysis and at least 2-fold variation in expressions, and we confine our attention to the top 5,000 genes with the highest sample variance. The 5,000 genes are standardized and partitioned into 100 subsets for DECO. The performance is assessed via 10-fold cross validation following the same approach in Section 5.1. The results are summarized in Table 4.3. As a reference, we also report these values for the null model.

Table 4.3: The results of all methods for mammalian eye diseases with \((n, p, m) = (120, 5000, 100)\)

<table>
<thead>
<tr>
<th>Method</th>
<th>MSE</th>
<th>Model size</th>
<th>runtime</th>
</tr>
</thead>
<tbody>
<tr>
<td>DECO-3</td>
<td>0.012</td>
<td>4.3</td>
<td>9.6</td>
</tr>
<tr>
<td>lasso-full</td>
<td>0.012</td>
<td>11</td>
<td>139.0</td>
</tr>
<tr>
<td>lasso-refine</td>
<td><strong>0.010</strong></td>
<td>11</td>
<td>139.7</td>
</tr>
<tr>
<td>lasso-naive</td>
<td>37.65</td>
<td>6.8</td>
<td>7.9</td>
</tr>
<tr>
<td>Null</td>
<td>0.021</td>
<td>—</td>
<td>—</td>
</tr>
</tbody>
</table>

Electricity load diagram  This dataset (Trindade, 2014) consists of electricity load from 2011 - 2014 for 370 clients. The data are originally recorded in KW for every 15 minutes, resulting in 14,025 attributes. Our goal is to predict the most recent electricity load by using all previous data points. The variance of the 14,025 features ranges from 0 to \(10^7\). To reduce the conditional number of the feature matrix, we remove features whose variances are below the lower 10% quantile (a value of \(10^5\)) and retain 126,231 features. We then expand the feature sets by including the interactions between the first 1,500 attributes that has the largest correlation with the clients’
most recent load. The resulting 1,251,980 features are then partitioned into 1,000 subsets for DECO. Because cross-validation is computationally demanding for such a large dataset, we put the first 200 clients in the training set and the remaining 170 clients in the testing set. We also scale the value of electricity load between 0 and 300, so that patterns are more visible. The results are summarized in Table 4.4.

Table 4.4: The results of all methods for electricity load diagram data with \((n, p, m) = (370, 1251980, 1000)\)

<table>
<thead>
<tr>
<th>Method</th>
<th>MSE</th>
<th>Model size</th>
<th>Runtime</th>
</tr>
</thead>
<tbody>
<tr>
<td>DECO-3</td>
<td>0.691</td>
<td>4</td>
<td>67.9</td>
</tr>
<tr>
<td>lasso-full</td>
<td>2.205</td>
<td>6</td>
<td>23,515.5</td>
</tr>
<tr>
<td>lasso-refine</td>
<td>1.790</td>
<td>6</td>
<td>22,260.9</td>
</tr>
<tr>
<td>lasso-naive</td>
<td>3.6 \times 10^8</td>
<td>4966</td>
<td>52.9</td>
</tr>
<tr>
<td>Null</td>
<td>520.6</td>
<td>—</td>
<td>—</td>
</tr>
</tbody>
</table>
5

DEME: A block-wise partitioning scheme

5.1 Introduction

In Chapter 3 and 4, we introduced two framework *message* and *DECO* in solving the "large \( n \)" and the "large \( p \)" problem. In particular, *message* algorithm partitions the data matrix row-wisely to down-scale the sample size while *DECO* partitions the data matrix column-wisely to break the high dimensionality. In many real applications, the datasets are often large in both the sample size and the dimension. Thus, each individual framework might be incapable to solve the large-scale problem. We address this issue in this chapter by developing a new "divide-and-conquer" framework *DEME* which can solve the "large-\( n \)-and-large-\( p \)" problem by leveraging both *message* and *DECO*. More precisely, it first applies *message* algorithm to partition the data matrix into row cubes and then applies *DECO* to partition each row cube into multiple small blocks. The feature selection algorithm will be running on these small blocks in parallel and the obtained results will be aggregated using *DECO* and *message* to output the final model. The concrete framework will be described in next section.
We remark that this new framework is extremely scalable as the data matrix can now be partitioned into small blocks with almost fixed sizes. This enables the possibility of building an easy expanding distributed system. In particular, any further increase in the dimension or the sample size can be solved by allocating the new data to extra workers on the row or column side without touching the existing system and data. This block-wise parallel feature also fits well in the GPU architecture. A simple analogy is that the CPU BLOCK’s correspond to the row cubes and the THREADs within each BLOCK process the small blocks within the associated row cubes.

This chapter is organized in two sections. In Section 5.1, we cover the detail on the DEME framework. Its theoretical property is then studied in Section 5.2.

### 5.2 The DEME framework

The structure of the DEME framework is simple: it just combines the message algorithm with the DECO framework after block partitioning the data. The concrete partitioning scheme is illustrated in Figure 5.1.

![Figure 5.1](image.png)

**Figure 5.1:** The partitioning scheme for DEME
DEME first partitions the data row-wisely into \( l \) row cubes, so that the sample size in each row cube is small. It then partitions each row cube column-wisely into \( m \) blocks to further reduce the dimensionality. The whole procedure is equivalently to partition the original data matrix into \( l \times m \) small blocks. The feature selection algorithm will be executed on each small block in parallel. After the models are selected on the \( l \times m \) blocks, DEME synthesizes these models in a reverse order, i.e., it will first combine the models produced by the blocks in the same row cube via \textit{DECO} to obtain \( l \) row-models. These row-models are then aggregated via \textit{message} to output the final model. The concrete algorithm is described in Algorithm 6.

**Algorithm 6** The DEME framework

**Initialization:**
1: Input \((Y, X), p, n, m, l\). Standardize \(X\) and \(Y\) to \(x\) and \(y\) with mean zero;

**Stage 1 : Block partitioning**
2: Partition (arbitrarily) \((y, x)\) into \( l \times m \) disjoint subsets \((y(i), x(i,j))\) with \(i \in [l], j \in [m]\);
3: Distribute the data into \( l \times m \) machines;

**Stage 2 : DECO estimation**
4: for \(i = 1\) to \(l\) do
5: Initialize \(F = 0\);
6: for \(j = 1\) to \(m\) do
7: \(F = F + x(i,j)x(i,j)^T\);
8: \(\hat{F} = \sqrt{p}(F + r_1I_p)^{-1/2}\) on the master machine and then pass back;
9: for \(j = 1\) to \(m\) do
10: \(\hat{y} = \hat{F}y(i)\) and \(\hat{x}^{(i,j)} = \hat{F}x^{(i,j)}\);
11: for \(j = 1\) to \(m\) do
12: Select features \(\hat{\gamma}^{(i,j)}\) either Algorithm 1 or 2;
13: \(\hat{\gamma}^{(i)} = (\hat{\gamma}^{(i,1)}, \hat{\gamma}^{(i,2)}, \ldots, \hat{\gamma}^{(i,m)} );\)
14: Do the refinement step if necessary.

**Stage 3 : Message synthesize**
15: Gather all subset models \(\hat{\gamma}^{(i)}\);
16: for \(k = 1\) to \(p\) do
17: \(\hat{\gamma}_k = \text{median}\{\hat{\gamma}_k^{(i)}, i = 1, 2, \ldots, l\};\)
18: return \(\hat{\gamma}\);

As shown in Algorithm 6, the DEME requires only one synchronizing step at Stage 2 for computing the decorrelation matrix. Thus, this framework can be well incorporated in the GPU computing architecture. In particular, one might distribute the \(l \times m\) blocks into \(l \times m\) GPU THREADs with the THREADs corresponding to
the same row cube located in the same GPU BLOCK. See Figure 5.2. In CUDA programming, THREADs within the same BLOCK are allowed to synchronize and thus the implementation of DEME is possible. Alternatively, the DEME framework well fits in the mapreduce framework using spark with the small blocks being the mapper’s and the row cube being the reducer. See Figure 5.2. Since the communication is limited to once for the entire algorithm, such implementation is expected to be efficient.
5.3 The consistency theory

The consistency of DEME can be directly implied from the consistency of message and DECO. Theory in Section 3.3 and 4.3 have justified the consistency of the two methods for both regularized regression and Bayesian variable selection. We synthesize the result as follows.

**Theorem 5.1** (DEME for regularized regression). Assume the conditions in Theorem 3.1 and 4.1 for Gaussian noise or Theorem 3.2 and 4.2 for heavy-tailed noise and the total sample size is \( N < p \). For each subset we choose \( \lambda = c \sigma \sqrt{\frac{\log p}{n_0}} \) where \( n_0 \) is the subset size. If \( c \) and \( n_0 \) satisfies the condition in Theorem 3.1 and ?? and \( n_0 \geq 116 \), then there exists some absolute constants \( K_1 > 0 \) such that

\[
P\left( \hat{\gamma} = \gamma_* \right) \geq 1 - \exp\left\{ - \frac{K_1 d_*^2}{32 c^2 M_0^2 \sigma^2 \log p} N \right\}.
\]

If \( c \) and \( n_0 \) satisfies the condition in Theorem 3.2 and \( n_0 \geq 328(\log p)^2 \), then there exists some absolute constants \( K_2 > 0 \) such that

\[
P\left( \hat{\gamma} = \gamma_* \right) \geq 1 - \exp\left\{ - \frac{K_2 d_*^4}{128 c^4 M_0^4 \sigma^4 (\log p)^2} N \right\}.
\]

**Theorem 5.2** (DEME for Bayesian variable selection). Assume the conditions in Theorem 3.4 and 4.3 for Gaussian noise or Theorem 3.5 and 4.4 for heavy-tailed noise and the total sample size is \( N < p \). For each subset we allocate \( n_0 \) samples and report the model within the sampled models that has the highest posterior probability. If \( \alpha, \kappa \) and \( n_0 \) satisfies the condition in Theorem 3.4 and 4.3 and also \( n_0 \geq 120 \), then there exists some absolute constants \( K'_1 > 0 \) such that

\[
P\left( \hat{\gamma} = \gamma_* \right) \geq 1 - \exp\left\{ - \frac{K'_1 d_*^2}{32(\alpha + \kappa) \sigma^2 M_0^2 \log p} N \right\}.
\]
If $\alpha, \kappa$ and $n_0$ satisfies the condition in Theorem 3.5 and $n_0 \geq 328(\log p)^2$, then there exists some absolute constants $K'_2 > 0$ such that

$$
P \left( \hat{\gamma} = \gamma_* \right) \geq 1 - \exp \left\{ - \frac{K'_2 d_* 2}{128(\alpha + \kappa)\sigma^2 M_0^2 (\log p)^2 N} \right\}.
$$

As shown in Theorem 5.1 and 5.2, DEME is consistent in feature selection and might even improve the accuracy when the noise is heavy-tailed. Considering the huge computation gain from the parallel computing, DEME is promising and scalable well for dataset with both large $n$ and large $p$. 
Conclusion and future extensions

In this thesis, I proposed two flexible and efficient frameworks for feature selection and estimation. The two frameworks are both embarrassingly parallel, eliminating the computational burden attributable to communication among machines, and are as efficient as other simple subset aggregation methods. The message algorithm partitions the dataset in the samples space and aggregate the subset model via majority voting. By selecting the median model, message can achieve better accuracy even than feature selection on the full data, resulting in an improvement also in MSE performance. Extensive simulation experiments show outstanding performance relative to competitors in terms of computation, feature selection and prediction. The DECO framework partitions the dataset in the feature space after applying a decorrelation step. The selected feature within each subsets are then concatenated to output the final model. DECO is shown to be theoretically attractive, empirically competitive and is straightforward to implement. In particular, I have shown that DECO achieves the same minimax convergence rate as if the full data were used and the rate does not depend on the number of partitions. I demonstrated the empirical performance of DECO via extensive experiments and compare it to various
approaches for fitting full data. As illustrated in the experiments, \textit{DECO} can not only reduce the computational cost substantially, but often outperform the full data approaches in terms of model selection and parameter estimation.

By leveraging both \textit{message} and \textit{DEME}, I propose a new framework \textit{DEME} that can deal with large-$p$-large-$n$ problems. More precisely, \textit{DEME} first partitions the large-$p$-large-$n$ dataset in the sample space to obtain $l$ row cubes such that each becomes a large-$p$-small-$n$ dataset. We then partition the feature space of each row cube into $m$ blocks. This procedure is equivalent to partitioning the original data matrix $X$ into $l \times m$ small blocks, each with a feasible size that can be stored and fitted in a computer. We then apply the \textit{DECO} framework to the small blocks in the same row cube using Algorithm 4. The last step is to apply Algorithm 3 to aggregate the $l$ row block estimators to output the final estimate. The whole procedure is extremely scalable.

The theory and the methodology in this thesis primarily focus on the linear model, part of which can be straightforwardly generalized to broader family of models, such as the generalized linear model. In particular, the theory of \textit{message} relies only on the Chernoff’s bound and a consistent feature selection on each subset, regardless of whether the model is linear or not. Thus, this framework can be easily extended to any feature selection algorithm that is proved to be consistent. \textit{DECO} is slightly more difficult to generalize, as the form of decorrelation might shift vastly for different types of models. One promising approach is to construct a more general preconditioning approach via Bayesian hierarchical prior and this remains as one of the most interesting future extensions to explore!
Appendix A

Theory on the elliptical family

In Appendix A, we provide the Theory on the elliptical family. In particular, we prove Theorem 2.4, 2.5 and 2.6.

A.1 Proof for Theorem 2.4

Define $\Phi = X^T(XX^T)^{-1}X$. It is easy to show Theorem 2.4 is equivalent to the following lemma.

**Lemma 4.** Assume $p > c_0n$ for some $c_0 > 1$, then for any $C > 0$ there exists some $0 < c_1 < 1 < c_2$ and $c_3 > 0$ such that for any $t > 0$ and any $i \in Q, j \neq i$, we have

$$P\left(|\Phi_{ii}| \leq \frac{c_1c_* n}{c^* p}\right) \leq 2e^{-Cn}, \quad P\left(|\Phi_{ii}| \geq \frac{c_2c_* n}{c^* p}\right) \leq 2e^{-Cn}$$

and

$$P\left(|\Phi_{ij}| \geq \frac{c_4c_* t \sqrt{n}}{c^* p}\right) \leq 5e^{-Cn} + 2e^{-t^2/2},$$

where $c_4 = \frac{\sqrt{c_2(c_0 - c_1)}}{\sqrt{c_3(c_0 - c)}}$ and $c_*, c^*$ are the smallest and the largest eigenvalues of $\Sigma$. In
addition, for any \( \gamma \in [p] \) and \( \beta \in \mathcal{R}^p \), we have

\[
P\left( |\beta^T \Phi_{ii} \beta| \leq \frac{c_1 c_p}{c_*} \|\Sigma_\gamma^2 \beta_\gamma\|_2^2 \frac{n}{p} \right) \leq 2e^{-Cn}, \quad P\left( |\beta^T \Phi_{ii} \beta| \geq \frac{c_2 c_p}{c_*} \|\Sigma_\gamma^2 \beta_\gamma\|_2^2 \frac{n}{p} \right) \leq 2e^{-Cn}
\]

To prove Lemma 4, we need following propositions and lemmas. We list all prerequisite results without proofs but refer readers to the materials for reference checking.

Let \( P \in \mathcal{O}(p) \) be a \( p \times p \) orthogonal matrix from the orthogonal group \( \mathcal{O}(p) \). Let \( H \) denote the first \( n \) columns of \( P \). Then \( H \) is in the Stiefel manifold (Chikuse, 2003). In general, the Stiefel manifold \( V_{n,p} \) is the space whose points are \( n \)-frames in \( \mathcal{R}^p \) represented as the set of \( p \times n \) matrices \( X \) such that \( X^T X = I_n \). Mathematically, we can write

\[
V_{n,p} = \{ X \in \mathcal{R}^{p \times n} : X^T X = I_n \}.
\]

There is a natural measure \((dX)\) called Haar measure on the Stiefel manifold, invariant under both right orthogonal and left orthogonal transformations. We standardize it to obtain a probability measure as \([dX] = (dX)/V(n,p)\), where \( V(n,p) = 2^n \pi^{np/2}/\Gamma_n(1/2p) \).

**Lemma 5.** (Chikuse, 2003, Page 41-44) Supposed that a \( p \times n \) random matrix \( Z \) has the density function of the form

\[
f_Z(Z) = |\Sigma|^{-n/2} g(Z^T \Sigma^{-1} Z),
\]

which is invariant under the right-orthogonal transformation of \( Z \), where \( \Sigma \) is a \( p \times p \) positive definite matrix. Then its orientation \( H_z = Z(Z^T Z)^{-1/2} \) has the matrix angular central Gaussian distribution (MACG) with a probability density function

\[
MACG(\Sigma) = |\Sigma|^{-n/2} |H_z^T \Sigma^{-1} H_z|^{-p/2}.
\]
In particular, if \( Z \) is a \( p \times n \) matrix whose distribution is invariant under both the left- and right-orthogonal transformations, then \( H_Y \), with \( Y = BZ \) for \( BB^T = \Sigma \), has the MACG(\( \Sigma \)) distribution.

When \( n = 1 \), the MACG distribution becomes the angular central Gaussian distribution, a description of the multivariate Gaussian distribution on the unit sphere (Watson et al., 1983).

**Lemma 6.** (Chikuse, 2003, Page 70, Decomposition of the Stiefel manifold) Let \( H \) be a \( p \times n \) random matrix on \( V_{n,p} \), and write

\[
H = (H_1 \ H_2),
\]

with \( H_1 \) being a \( p \times q \) matrix where \( 0 < q < n \). Then we can write

\[
H_2 = G(H_1)U_1,
\]

where \( G(H_1) \) is any matrix chosen so that \( (H_1 \ G(H_1)) \in O(p) \); as \( H_2 \) runs over \( V_{n-q,p} \), \( U_1 \) runs over \( V_{n-q,p-q} \) and the relationship is one to one. The differential form \([dH]\) for the normalized invariant measure on \( V_{n,p} \) is decomposed as the product

\[
[dH] = [dH_1][dU_1]
\]

of those \([dH_1]\) and \([dU_1]\) on \( V_{q,p} \) and \( V_{n-q,p-q} \), respectively.

**Lemma 7.** ([Lemma 4 in Fan and Lv (2008)]) Let \( U \) be uniformly distributed on the Stiefel manifold \( V_{n,p} \). Then for any \( C > 0 \), there exist \( c'_1, c'_2 \) with \( 0 < c'_1 < 1 < c'_2 \), such that

\[
P\left( e_1^TUU^Te_1 < c'_1 \frac{n}{p} \right) \leq 2e^{-Cn},
\]

and

\[
P\left( e_1^TUU^Te_1 > c'_2 \frac{n}{p} \right) \leq 2e^{-Cn}.
\]
Proposition 1. Let $\chi^2(n)$ denote the chi-square random variable with degree of freedom $n$. For any $t > 0$ we have

$$\mathbb{P}\left(\frac{1}{n} \chi^2(n) \geq 1 + 2\sqrt{\frac{t}{n} + \frac{2t}{n}}\right) \leq e^{-t} \quad \text{and} \quad \mathbb{P}\left(\frac{1}{n} \chi^2(n) \leq 1 - 2\sqrt{\frac{t}{n}}\right) \leq e^{-t}. \quad \text{(A.1)}$$

Thus, for any $C > 0$, there exists some $c_3 > 0$ such that

$$\mathbb{P}\left(\frac{1}{n} \chi^2(n) \leq c_3\right) \leq e^{-Cn}.$$ 

Now we prove the main results.

Proof of Lemma 4. According to Lemma 1, we consider a transformed $n \times p$ random matrix $Z = \bar{L}^{-1}X\Sigma^{-1/2}$, where $\bar{L} = \text{diag}(\sqrt{p}L_1/\|z_1\|_2, \ldots, \sqrt{p}L_n/\|z_n\|_2)$. By definition, $Z$ follows standard multivariate Gaussian. Consider its SVD decomposition,

$$Z = VDU^T,$$

where $V \in O(n)$, $D$ is a diagonal matrix and $U$ is a $p \times n$ random matrix belonging to the Stiefel manifold $V_{n,p}$. With such notion, we can rewrite the projection matrix as

$$X^T(XXT)^{-1}X = \Sigma^{1/2}UL(U^T\Sigma U)^{-1}U^T\Sigma^{1/2} = HH^T,$$

where $H = \Sigma^{1/2}U(U^T\Sigma U)^{-1/2}$ and $H \in V_{n,p-1}$. Therefore, the two quantities that we are interested in are $\Phi_{ii} = e_i^T H H^T e_i$ (diagonal term) and $\Phi_{ij} = e_i^T H H^T e_j$ (off-diagonal term), where $e_i^T$ is the $p-$dimensional unit vector with the $i^{th}$ coordinate being one. The proof is divided into two parts, where in the first part we consider diagonal terms and the second part takes care of off-diagonal terms.

Part I: First, we consider the diagonal term $e_i^T H H^T e_i$. Recall the definition of $H$ and

$$e_i^T H H^T e_i = e_i^T \Sigma^{1/2}U(U^T\Sigma U)^{-1}U^T\Sigma^{1/2}e_i.$$
There always exists some orthogonal matrix $Q$ that rotates the vector $\Sigma^{\frac{1}{2}}e_i$ to the direction of $e_1$, i.e,

$$\Sigma^{\frac{1}{2}}v = \|\Sigma^{\frac{1}{2}}v\|Qe_1.$$ 

Then we have

$$e_i^T HH^T e_i = \|\Sigma^{\frac{1}{2}}e_i\|^2 e_i^T Q^T U(U^T \Sigma U)^{-1} U^T Q e_1 = \|\Sigma^{\frac{1}{2}}v\|^2 e_1^T \tilde{U}(U^T \Sigma U)^{-1} \tilde{U} e_1,$$

where $\tilde{U} = Q^T U$ is uniformly distributed on $V_{n,p}$, because $U$ is uniformly distributed on $V_{n,p}$ (see discussion in the beginning). Now the magnitude of $e_i^T HH^T e_i$ can be evaluated in two parts. For the norm of the vector $\Sigma^{\frac{1}{2}}v$, we have

$$\lambda_{\min}(\Sigma) \leq e_i^T \Sigma e_i = \|\Sigma^{\frac{1}{2}}e_i\|^2 \leq \lambda_{\max}(\Sigma), \quad (A.2)$$

and for the remaining part,

$$e_1^T \tilde{U}(U^T \Sigma U)^{-1} \tilde{U} e_1 \leq \lambda_{\max}((U^T \Sigma U)^{-1}) \|\tilde{U} e_1\|^2 \leq \lambda_{\min}(\Sigma)^{-1} \|\tilde{U} e_1\|^2,$$

and

$$e_1^T \tilde{U}(U^T \Sigma U)^{-1} \tilde{U} e_1 \geq \lambda_{\min}((U^T \Sigma U)^{-1}) \|\tilde{U} e_1\|^2 \geq \lambda_{\max}(\Sigma)^{-1} \|\tilde{U} e_1\|^2.$$

Consequently, we have

$$e_i^T HH^T e_i \leq \lambda_{\max}(\Sigma) \lambda_{\min}(\Sigma)^{-1} e_i^T UU^T e_1, \quad e_i^T HH^T e_i \geq \lambda_{\min}(\Sigma) \lambda_{\max}(\Sigma)^{-1} e_i^T UU^T e_1. \quad (A.3)$$

Therefore, following Proposition 7, for any $C > 0$ we have

$$P\left( e_i^T HH^T e_i \leq \frac{c_1 c_* n}{c^* p} \right) \leq 2e^{-Cn},$$

and

$$P\left( e_i^T HH^T e_i \geq \frac{c_2 c_* n^1}{c^* p} \right) \leq 2e^{-Cn}.$$
Denoting $c_1'$ by $c_1$ and $c_2'$ by $c_2$, we obtain the first equation in Lemma 4. In addition, if we replace $e_i$ by $\beta_\gamma$ in the above proof, everything will follow the same, thus we have

$$P\left(\beta_\gamma^T H H^T \beta_\gamma \leq \frac{c_1 c_* n}{c_* p}\right) \leq 2e^{-Cn},$$

and

$$P\left(\beta_\gamma^T H H^T \beta_\gamma \geq \frac{c_2 c_* n^1}{c_* p}\right) \leq 2e^{-Cn}.$$

which is the last equation in Lemma 4.

**Part II:** Second, for off-diagonal terms, the proof is almost identical to the proof of Lemma 5 in Wang and Leng (2015).

The proof depends on the decomposition of Stiefel manifold. Without loss of generality, we prove the bound only for $e_2^T H H^T e_1$, then the other off-diagonal terms should follow exactly the same argument. According to Lemma 6, we can decompose $H = (T_1, H_2)$ with $T_1 = G(H_2) H_1$, where $H_2$ is a $p \times (n - 1)$ matrix, $H_1$ is a $(p - n + 1) \times 1$ vector and $G(H_2)$ is a matrix such that $(G(H_2), H_2) \in O(p)$. The invariant measure on the Stiefel manifold can be decomposed as

$$[H] = [H_1][H_2]$$

where $[H_1]$ and $[H_2]$ are Haar measures on $V_{1,n-p+1}, V_{n-1,p}$ (Notice that $q = n - 1$ in this decomposition) respectively. As pointed out before, $H$ has the $MACG(\Sigma)$ distribution, which possesses a density as

$$p(H)\propto |H^T \Sigma^{-1} H|^{-p/2}[dH].$$

Using the identity for matrix determinant

$$\begin{vmatrix} A & B \\ C & D \end{vmatrix} = |A||D - CA^{-1}B| = |D||A - BD^{-1}C|,$$
we have

\[
P(H_1, H_2) \propto |H_2^T \Sigma^{-1} H_2|^{-\nu/2} (T_1^T \Sigma^{-1} T_1 - T_1^T \Sigma^{-1} H_2 (H_2^T \Sigma^{-1} H_2)^{-1} H_2^T \Sigma^{-1} T_1)^{-\nu/2}
\]

\[
= |H_2^T \Sigma^{-1} H_2|^{-\nu/2} (H_2^T G(H_2)^T (\Sigma^{-1} - \Sigma^{-1} H_2 (H_2^T \Sigma^{-1} H_2)^{-1} H_2^T \Sigma^{-1}) G(H_2) H_1)^{-\nu/2}
\]

\[
= |H_2^T \Sigma^{-1} H_2|^{-\nu/2} (H_2^T G(H_2)^T \Sigma^{-1/2} (I - T_2) \Sigma^{-1/2} G(H_2) H_1)^{-\nu/2},
\]

where \( T_2 = \Sigma^{-1/2} H_2 (H_2^T \Sigma^{-1} H_2)^{-1} H_2^T \Sigma^{-1/2} \) is an orthogonal projection onto the linear space spanned by the columns of \( \Sigma^{-1/2} H_2 \). It is easy to verify the following result by using the definition of \( G(H_2) \),

\[
[\Sigma^{1/2} G(H_2)(G(H_2)^T \Sigma G(H_2))^{-1/2}, \Sigma^{-1/2} H_2 (H_2^T \Sigma^{-1} H_2)^{-1/2}] \in \mathcal{O}(p),
\]

and therefore we have

\[
I - T_2 = \Sigma^{1/2} G(H_2)(G(H_2)^T \Sigma G(H_2))^{-1} G(H_2)^T \Sigma^{1/2},
\]

which simplifies the density function as

\[
P(H_1, H_2) \propto |H_2^T \Sigma^{-1} H_2|^{-\nu/2} (H_2^T (G(H_2)^T \Sigma G(H_2))^{-1} H_1)^{-\nu/2}.
\]

Now it becomes clear that \( H_1 | H_2 \) follows the Angular Central Gaussian distribution \( ACG(\Sigma') \), where

\[
\Sigma' = G(H_2)^T \Sigma G(H_2).
\]

Next, we relate the target quantity \( e_1^T H H^T e_2 \) to the distribution of \( H_1 \). Notice that for any orthogonal matrix \( Q \in \mathcal{O}(n) \), we have

\[
e_1^T H H^T e_2 = e_1^T H Q Q^T H^T e_2 = e_1^T H' H'^T e_2.
\]

Write \( H' = HQ = (T_1', H_2') \), where \( T_1' = [T_1^{(1)}, T_1^{(2)}, \cdots, T_1^{(p)}] \), \( H_2' = [H_2^{(i,j)}] \). If we choose \( Q \) such that the first row of \( H_2' \) are all zero (this is possible as we can choose the first column of \( Q \) being the first row of \( H \) upon normalizing), i.e.,

\[
e_1^T H' = [T_1^{(1)}, 0, \cdots, 0] \quad e_2^T H' = [T_1^{(2)}, H_2^{(2,1)}, \cdots, H_2^{(2,n-1)}],
\]

82
then immediately we have \( e_1^T H H^T e_2 = e_1^T H' H'^T e_2 = T_1^{(1)} T_1^{(2)} \). This indicates that

\[
e_1^T H H^T e_2 \overset{(d)}{=} T_1^{(1)} T_1^{(2)} \bigg| e_1^T H_2 = 0.
\]

As shown at the beginning, \( H_1 \) follows \( ACG(\Sigma') \) conditional on \( H_2 \). Let \( H_1 = (h_1, h_2, \cdots, h_p)^T \) and let \( x^T = (x_1, x_2, \cdots, x_{p-n+1}) \sim N(0, \Sigma') \), then we have

\[
h_i \overset{(d)}{=} \frac{x_i}{\sqrt{x_1^2 + \cdots + x_{p-n+1}^2}}.
\]

Notice that \( T_1 = G(H_2)H_1 \), a linear transformation on \( H_1 \). Defining \( y = G(H_2)x \), we have

\[
T_1^{(i)} \overset{(d)}{=} \frac{y_i}{\sqrt{y_1^2 + \cdots + y_p^2}},
\]

where \( y \sim N(0, G(H)\Sigma' G(H)^T) \) is a degenerate Gaussian distribution. This degenerate distribution contains an interesting form. Letting \( z \sim N(0, \Sigma) \), we know \( y \) can be expressed as \( y = G(H)G(H)^T z \). Write \( G(H_2)^T \) as \([g_1, g_2]\) where \( g_1 \) is a \((p-n+1) \times 1\) vector and \( g_2 \) is a \((p-n+1) \times (p-1)\) matrix, then we have

\[
G(H_2)G(H_2)^T = \begin{pmatrix} g_1^T g_1 & g_1^T g_2 \\ g_2^T g_1 & g_2^T g_2 \end{pmatrix}.
\]

We can also write \( H_2^T = [0_{n-1,1}, h_2] \) where \( h_2 \) is a \((n-1) \times (p-1)\) matrix, and using the orthogonality, i.e., \([H_2 \ G(H_2)] \ [H_2 \ G(H_2)]^T = I_p\), we have

\[
g_1^T g_1 = 1, \ g_1^T g_2 = 0_{1,p-1} \quad \text{and} \quad g_2^T g_2 = I_{p-1} - h_2 h_2^T.
\]

Because \( h_2 \) is a set of orthogonal basis in the \( p-1 \) dimensional space, \( g_2^T g_2 \) is therefore an orthogonal projection onto the space \( \{h_2\}^\perp \) and \( g_2^T g_2 = AA^T \) where \( A = g_2^T (g_2 g_2^T)^{-1/2} \) is a \((p-1) \times (p-n)\) orientation matrix on \( \{h_2\}^\perp \). Together, we have

\[
y = \begin{pmatrix} 1 & 0 \\ 0 & AA^T \end{pmatrix} z.
\]
This relationship allows us to marginalize $y_1$ out with $y$ following a degenerate Gaussian distribution.

We now turn to transform the condition $e_1^T H_2 = 0$ onto constraints on the distribution of $T_1^{(i)}$. Letting $t_i^2 = e_1^T H H^T e_1$, then $e_1^T H_2 = 0$ is equivalent to $T_1^{(1)} = e_1^T H H^T e_1 = t_i^2$, which implies that

$$e_1^T H H^T e_2 \overset{(d)}{=} T_1^{(1)} T_1^{(2)} \bigg| T_1^{(1)^2} = e_1^T H H^T e_1.$$ 

Because the magnitude of $e_1^T H H^T e_1$ has been obtained in Part I, we can now condition on the value of $e_1^T H H^T e_1$ to obtain the bound on $T_1^{(2)}$. From $T_1^{(1)^2} = t_i^2$, we obtain that,

$$\left(1 - t_i^2\right)y_1^2 = t_i^2(y_2^2 + y_3^2 + \cdots + y_p^2).$$

(A.5)

Notice this constraint is imposed on the norm of $\tilde{y} = (y_2, y_3, \cdots, y_p)$ and is thus independent of $(y_2/\lVert \tilde{y} \rVert, \cdots, y_p/\lVert \tilde{y} \rVert)$. Equation (A.5) also implies that

$$(1 - t_1^2)(y_1^2 + y_2^2 + \cdots + y_p^2) = y_2^2 + y_3^2 + \cdots + y_p^2.$$ 

(A.6)

Therefore, combining (A.4) with (A.5), (A.6) and integrating $y_1$ out, we have

$$T_1^{(i)} \bigg| T_1^{(1)} = t_1 \overset{(d)}{=} \frac{\sqrt{1 - t_i^2}y_i}{\sqrt{y_2^2 + \cdots + y_p^2}}, \quad i = 2, 3, \cdots, p,$$

where $(y_2, y_3, \cdots, y_p) \sim N(0, AA^T \Sigma_{22} AA^T)$ with $\Sigma_{22}$ being the covariance matrix of $z_2, \cdots, z_p$.

To bound the numerator, we use the classical tail bound on the normal distribution as for any $t > 0$, ($\sigma_i = \sqrt{\text{var}(y_i)} \leq \sqrt{\lambda_{\max}(AA^T \Sigma_{22} AA^T)} \leq \lambda_{\max}(\Sigma)^{1/2}$),

$$P(|y_i| > t\sigma_i) = P(|y_i| > t\lambda_{\max}^{1/2}(\Sigma)) \leq 2e^{-t^2/2}.$$ 

(A.7)

For the denominator, letting $\tilde{z} \sim N(0, I_{p-1})$, we have

$$\tilde{y} = AA^T \Sigma_{22}^{1/2} \tilde{z} \quad \text{and} \quad \tilde{y}^T \tilde{y} = \tilde{z}^T \Sigma_{22}^{1/2} AA^T \Sigma_{22}^{1/2} \tilde{z} \overset{(d)}{=} \sum_{i=1}^{p-n} \lambda_i \tilde{x}_i^2(1),$$

84
where $X_i^2(1)$ are i.i.d chi-square random variables and $\lambda_i$ are non-zero eigenvalues of matrix $\Sigma_{22}^{1/2} A A^T \Sigma_{22}^{1/2}$. Here $\lambda_i$'s are naturally upper bounded by $\lambda_{\text{max}}(\Sigma)$.

To give lower bound the denominator, we notice that $\Sigma_{22}^{1/2} A A^T \Sigma_{22}^{1/2}$ and $A \Sigma_{22} A^T$ possess the same set of non-zero eigenvalues, thus

$$\min_i \lambda_i \geq \lambda_{\text{min}}(A \Sigma_{22} A^T) \geq \lambda_{\text{min}}(\Sigma).$$

Therefore,

$$\lambda_{\text{min}}(\Sigma) \sum_{i=1}^{p-n} X_i^2(1) \frac{p-n}{p-n} \leq \tilde{y}^T \tilde{y} \leq \lambda_{\text{max}}(\Sigma) \sum_{i=1}^{p-n} X_i^2(1) \frac{p-n}{p-n}.$$

The quantity $\frac{\sum_{i=1}^{p-n} X_i^2(1)}{p-n}$ can be bounded by Proposition 1. Thus, we have for any $C > 0$, there exists some $c_3 > 0$ such that

$$P\left(\tilde{y}^T \tilde{y} / (p-n) < c_3 \lambda_{\text{min}}(\Sigma)\right) \leq e^{-C(p-n)}.$$

Therefore, $T_1^{(2)}$ can be bounded as

$$P\left(|T_1^{(2)}| > \frac{\sqrt{c^*} \sqrt{1 - t_1^2}}{\sqrt{c_3} c_3 \sqrt{p-n}} |T_1^{(1)} = t_1\right) \leq e^{-C(p-n)} + 2e^{-t^2/2}.$$

Using the results from the diagonal term, we have

$$P\left(t_1^2 > \frac{c_2 c^* n}{c^*_*} \frac{n}{p}\right) \leq 2e^{-Cn} \quad \text{and} \quad P\left(t_1^2 < \frac{c_1 c^* n}{c^*_*} \frac{n}{p}\right) \leq 2e^{-Cn}.$$

Consequently, we have

$$P\left(|e_1^T \Sigma_{22} e_2| > \frac{c_4 c^* t \sqrt{n}}{c^*_*} \frac{p}{p}\right) = P\left(|T_1^{(1)} T_1^{(2)}| > \frac{c_4 c^* t \sqrt{n}}{c^*_*} \frac{n}{p} |T_1^{(1)} = t_1\right)$$

$$\leq P\left(|T_1^{(1)}|^2 > \frac{c_2 c^* n}{c^*_*} \frac{p}{p} |T_1^{(1)} = t_1\right) + P\left(|T_1^{(2)}| > \frac{\sqrt{c^*} t \sqrt{1 - c_1 n/p}}{\sqrt{c_3} c_3 \sqrt{p-n}} |T_1^{(1)} = t_1\right)$$

$$\leq 5e^{-Cn} + 2e^{-t^2/2},$$

where $c_4 = \frac{\sqrt{c_2 (c_0 - 1)}}{\sqrt{c_3 (c_0 - 1)}}$. \qed
A.2 Proofs for Theorem 2.5 and 2.6

We need following three propositions for our proof.

**Proposition 2.** (Lounici, 2008; Nemirovski, 2000; Akritas et al., 2014) Let $Y_i \in \mathcal{R}^p$ be random vectors with zero means and finite variances. Then we have for any $k$ norm with $k \in [2, \infty]$ and $p \geq 3$, we have

$$E \left\| \sum_{i=1}^{n} Y_i \right\|^2_k \leq \tilde{C} \min \{ k, \log p \} \sum_{i=1}^{n} E \| Y_i \|^2_k,$$

(A.8)

where $\tilde{C}$ is some absolute constant.

**Proposition 3.** Let $\chi^2(n)$ denote the chi-square random variable with degree of freedom $n$. For any $t > 0$ we have

$$\mathbb{P}\left( \frac{1}{n} \chi^2(n) \geq 1 + 2 \sqrt{\frac{t}{n} + \frac{2t}{n}} \right) \leq e^{-t} \quad \text{and} \quad \mathbb{P}\left( \frac{1}{n} \chi^2(n) \leq 1 - 2 \sqrt{\frac{t}{n}} \right) \leq e^{-t}. \quad (A.9)$$

**Proposition 4.** Let $Z \sim N(0, I_p)$, then we have the minimum eigenvalue of $ZZ^T/p$ satisfies that

$$\mathbb{P}\left\{ \lambda_{\min}(ZZ^T) \geq (\sqrt{p} - \sqrt{n} - \sqrt{t})^2, \lambda_{\max}(ZZ^T) \leq (\sqrt{p} + \sqrt{n} + \sqrt{t})^2 \right\} \geq 1 - 2 \exp(-t/2)$$

for any $t > 0$. Assume $p > c_0 n$ for $c_0 > 1$ and take $t = 4 \log p$, then if $p/\log p \geq 4/v_1$, we have

$$\mathbb{P}\left\{ \lambda_{\min}(ZZ^T/p) \geq v_1, \lambda_{\max}(ZZ^T/p) \leq v_2 \right\} \geq 1 - 2p^{-2} \quad (A.10)$$

where $v_1 = (1 - 2c_0)^2, v_2 = (1 + 2c_0)^2.\c_0$

The proof follows Corollary 5.35 in Vershynin (2010).

**Theorem A.1** (Concentration with heavy-tailed errors). Assume $x_i \sim EN(L, \Sigma)$ with $p > c_0 n$ and $\text{var}(\varepsilon) = \sigma^2$. Define $\tilde{X} = (XX^T/p)^{-\frac{1}{2}} X$ and $\tilde{\varepsilon} = (XX^T/p)^{-\frac{1}{2}} \varepsilon$. 

86
If \( \mathbb{E} \left[ \frac{P}{S_{x_{1}}} \right] = \mathbb{E} [L^{-2}] = l_{2} \) and \( p / \log p \geq 16c_{0} / (\sqrt{c_{0}} - 1)^{2} \), then for any \( t > 0 \) and \( r > 0 \), we have

\[
P \left( \frac{1}{n} \| \tilde{X}^{T} \tilde{\xi} \|_{\infty} \geq \frac{r}{\sqrt{n}} \right) \leq C_{1} \frac{t^{2} \sigma^{2} \log p}{r^{2}} + 2npe^{-\frac{t^{2}}{2}} + 2pe^{-Cn} + 3p^{-3},
\]

where \( C_{1} = \frac{14c_{2}c^{2}l^{2}}{v_{1}c^{2}_{0}} \) and \( c_{2}, c_{*}, c^{*} \) are defined in Theorem 2.4 and \( v_{1} \) defined in Lemma 4. If it is further true that \( \mathbb{E} [\xi_{i}^{4} / L_{i}^{4}] \leq M_{4} \) for some \( M_{4} > 0 \), then we have

\[
P \left( \frac{1}{4v_{2}} l_{2} \sigma^{2} \leq \frac{1}{n} \| \tilde{\xi} \|_{2}^{2} \leq \frac{4}{v_{1}} l_{2} \sigma^{2} \right) \geq 1 - \frac{M_{4}}{n} - 4p^{-2}.
\]

**Proof of Theorem A.1.** Recall the representation \( X = \tilde{L}Z\Sigma^{1/2} \) where

\[\tilde{L} = \text{diag}(\sqrt{p}L_{1}/\|z_{1}\|_{2}, \sqrt{p}L_{2}/\|z_{2}\|_{2}, \cdots, \sqrt{p}L_{n}/\|z_{n}\|_{2})\]

and \( Z = (z_{1}, z_{2}, \cdots, z_{n})^{T} \) is an \( n \times p \) matrix consists of independent Gaussian entries, i.e., \( z_{i} \sim N(0, I_{p}) \).

Define \( \eta = \tilde{X}^{T} \tilde{\xi} = X^{T}(XX^{T}/p)^{-1} \xi \) and \( \xi = \tilde{\xi}^{T} \tilde{\xi} = \tilde{\xi}^{T}(XX^{T})^{-1} \tilde{\xi} \). Let \( A = pX^{T}(XX^{T})^{-1} \tilde{L} \), then \( \eta = AL^{-1} \xi \).

**Part 1. Bounding \(|A_{ij}|\).** Consider the standard SVD on \( Z \) as \( Z = VDU^{T} \), where \( V \) and \( D \) are \( n \times n \) matrices and \( U \) is a \( p \times n \) matrix. Because \( Z \) is a matrix of iid Gaussian variables, its distribution is invariant under both left and right orthogonal transformation. In particular, for any \( T \in \mathcal{O}(n) \), we have

\[TVDU^{T} \stackrel{(d)}{=} VDU^{T},\]

i.e., \( V \) is uniformly distributed on \( \mathcal{O}(n) \) conditional on \( U \) and \( D \) (they are in fact independent, but we don’t need such a strong condition). Therefore, we have

\[
A = pX^{T}(XX^{T})^{-1} \tilde{L} = p\Sigma^{1/2}Z^{T}(\tilde{L}Z\Sigma Z^{T}\tilde{L})^{-1} \tilde{L} = p\Sigma^{1/2}UDV^{T} \tilde{L}(LVDU^{T}\Sigma UDV^{T} \tilde{L})^{-1} \tilde{L} = p\Sigma^{1/2}U(U^{T}\Sigma U)^{-1}D^{-1}V^{T} = \sqrt{p}\Sigma^{1/2}U(U^{T}\Sigma U)^{-1}(\frac{D}{\sqrt{p}})^{-1}V^{T}.
\]

87
Because \( V \) is uniformly distributed conditional on \( U \) and \( D \), the distribution of \( A \) is also invariant under right orthogonal transformation conditional on \( U \) and \( D \), i.e., for any \( T \in \mathcal{O}(n) \), we have

\[
A^{(d)} = AT. \tag{A.11}
\]

Our first goal is to bound the magnitude of individual entries \( A_{ij} \). Let \( v_i = e_i^T A A^T e_i \), which is a function of \( U \) and \( D \) (see below). From (A.11), we know that \( e_i^T A \) is uniformly distributed on the sphere \( S^{n-1}(\sqrt{v_i}) \) if conditional on \( v_i \) (i.e., conditional on \( U, D \)), which implies that

\[
e_i^T A = \sqrt{v_i} \left( \frac{x_1}{\sqrt{\sum_{j=1}^{n} x_j^2}}, \frac{x_2}{\sqrt{\sum_{j=1}^{n} x_j^2}}, \cdots, \frac{x_n}{\sqrt{\sum_{j=1}^{n} x_j^2}} \right), \tag{A.12}
\]

where \( x_j's \) are iid standard Gaussian variables. Thus, \( A_{ij} \) can be bounded easily if we can bound \( v_i \). Notice that for \( v_i \) we have

\[
v_i = e_i^T A A^T e_i = pe_i^T \Sigma^{1/2} U (U^T \Sigma U)^{-1} \left( \frac{D^2}{p} \right)^{-1} U^T \Sigma^{1/2} \Sigma e_i.
\]

\[
= pe_i^T H (U^T \Sigma U)^{-1/2} \left( \frac{D^2}{p} \right)^{-1} (U^T \Sigma U)^{-1/2} H^T e_i
\]

\[
\leq pe_i^T H H^T e_i \cdot \lambda_{\min}^{-1} (U^T \Sigma U) \cdot \lambda_{\min}^{-1} \left( \frac{D^2}{p} \right)
\]

Here \( H = \Sigma^{1/2} U (U^T \Sigma U)^{-1/2} \) is defined the same as in Wang and Leng (2015) and [cite] can be bounded as \( e_i^T H H^T e_i \leq \frac{c_{2\hat{c}^*}^*}{c_{2^*}} \) with probability \( 1 - 2 \exp(-Cn) \) (see the proof of Lemma 3 in Wang et al. (2015b)). Therefore, we have

\[
P \left( v_i \leq \frac{c_{2\hat{c}^*}^*}{c_{2^*}} \lambda_{\min}^{-1} \left( \frac{D^2}{p} \right) n \right) \geq 1 - 2 \exp(-Cn)
\]

Now applying the tail bound and the concentration inequality of the Gaussian and the \( X^2 \) variables to (A.12) we have for any \( t > 0 \) and any \( n > 8 \log p \)

\[
P(|v_j| > t) \leq 2 \exp(-t^2/2) \quad P \left( \frac{\sum_{j=1}^{n} x_j^2}{n} \leq \frac{1}{2} \right) \leq p^{-4}. \tag{A.13}
\]
Putting the pieces all together, we have for any $t > 0$ and any $C > 0$ that

$$
P\left( \max_{ij} |A_{ij}| \leq t \frac{c_* \sqrt{2e^2}}{c_*} \lambda_{\min}^{-\frac{1}{2}} \left( \frac{D^2}{p} \right) \right) \geq 1 - 2np \exp(-t^2/2) - 2p \exp(-Cn) - p^{-3}.
$$

Now according to (A.10), we can further bound $\lambda_{\min}(D^2/p)$ and obtain that

$$
P\left( \max_{ij} |A_{ij}| \leq \sqrt{\frac{2c_* c^2}{c_*^2 v_1}} \right) \geq 1 - 2np \exp(-t^2/2) - 3p \exp(-Cn) - 3p^{-3}. \quad (A.14)
$$

**Part 2. Bounding $\eta$**

The second step is to use (A.14) and Proposition 2 to bound $\eta$. The procedure follows similarly as in Lounici’s paper. We first note that $\|z_i\|^2$ follows a chi-square distribution $\chi^2(p)$. We have for any $t$

$$
P\left( \frac{\|z_i\|^2}{p} \geq 1 + 2 \sqrt{\frac{t}{p} + \frac{2t}{p}} \right) \leq e^{-t},
$$

from which we know

$$
P\left( \max_i p^{-1}\|z_i\|^2 < 5/2 \right) \geq 1 - pe^{-p/4}. \quad (A.15)
$$

Now define $W_j = (A_1 p^{-1/2}\|z_j\|_2 L_j^{-1} \varepsilon_j, A_2 p^{-1/2}\|z_j\|_2 L_j^{-1} \varepsilon_j, \ldots, A_p p^{-1/2}\|z_j\|_2 L_j^{-1} \varepsilon_j)$. It’s clear that $\eta = \sum_{j=1}^n W_j$. Applying Proposition 2 to $W_j$’s with the $l_\infty$ norm and noticing that $L_j$ is independent of $z_j$ we have

$$
E\left\| \sum_{j=1}^n W_j \right\|_\infty^2 \leq \log p \sum_{j=1}^n E\left| W_j \right|_\infty^2 \leq \log p \frac{14c_* c^2}{c_*^2 v_1} \sigma^2 \gamma_2^2 \sum_{j=1}^n E[L_j^{-2}] \leq \frac{14c_* c^2}{c_*^2 v_1} \sigma^2 \gamma_2^2 \gamma_2 n \log p.
$$

Using the Markov inequality on $\eta$, we have for any $r > 0$

$$
P\left( \frac{1}{n} \|\eta\|_\infty \geq \frac{r}{\sqrt{n}} \right) = P\left( \frac{1}{\sqrt{n}} \|\eta\|_\infty \geq r \right) \leq \frac{E\|\eta\|_\infty^2}{nr^2} = \frac{E\|\sum_{j=1}^n W_j \|_\infty^2}{nr^2} \leq \frac{14c_* c^2 \sigma^2 \gamma_2^2 \gamma_2^2}{c_*^2 v_1 r^2}.
$$
Part 3. Bound $\xi$ Define $\tilde{L} = \text{diag}(L_1, \ldots, L_n)$ and $\tilde{Z} = p^{-1/2} \text{diag}(\|z_1\|_2, \ldots, \|z_n\|_2)$, then we have

$$\xi = \varepsilon^T L^{-T} \tilde{Z}^T (Z\Sigma Z^T / p)^{-1} \tilde{Z} L^{-1} \varepsilon.$$ 

Thus, we know

$$\min_i \|z_i\|_2^2 \lambda^{-1} \|Z\Sigma Z^T / p\| L^{-1} \varepsilon \leq \xi \leq \max_i \|z_i\|_2^2 \lambda^{-1} \|Z\Sigma Z^T / p\| L \varepsilon.$$ 

Using the tail bound on $\mathcal{X}^2(p)$ we have

$$\mathbb{P}\left( \max_i p^{-1} \|z_i\|_2^2 < 5/2, \ \min_i p^{-1} \|z_i\|_2^2 > 1/2 \right) \geq 1 - 2pe^{-p/4}. \quad (A.16)$$

and Proposition 4 shows that

$$\mathbb{P}\left\{ \lambda_{\min}(ZZ^T / p) \geq v_1, \ \lambda_{\max}(ZZ^T / p) \leq v_2 \right\} \geq 1 - 2p^{-2}$$

Thus, we just need to bound $Q = \|L^{-1} \varepsilon\|_2^2$. First, we have $\mathbb{E}[Q] = nl_2\sigma^2$ and also

$$\mathbb{E}[Q - \mathbb{E}[Q]]^2 = \sum_{i=1}^{n} \mathbb{E}[(L_i^{-2} \varepsilon_i^2)^2 - l_2\sigma^2]^2 + \sum_{i \neq j} \mathbb{E}[(L_i^{-2} \varepsilon_i^2 - l_2\sigma^2)(L_j^{-2} \varepsilon_j^2 - l_2\sigma^2)]$$

$$= \sum_{i=1}^{n} \mathbb{E}[(L_i^{-2} \varepsilon_i^2 - l_2\sigma^2)^2] \leq \sum_{i=1}^{n} \mathbb{E}[L_i^{-2} \varepsilon_i^2] \leq nM_4.$$ 

Now using the Chebyshev’s inequality, we have

$$\mathbb{P}(\|Q - nl_2\sigma^2\| > t) \leq \frac{nM_4}{t^2}.$$ 

Taking $t = nl_2\sigma^2/2$ we get

$$\mathbb{P}\left( \frac{1}{2} nl_2\sigma^2 \leq \|L^{-1} \varepsilon\|_2^2 \leq \frac{3}{2} nl_2\sigma^2 \right) \leq \frac{M_4}{n}.$$ 

Putting all pieces together, we obtain that

$$\mathbb{P}\left( \frac{1}{4v_2} l_2\sigma^2 \leq \frac{1}{n} \xi \leq \frac{4}{v_1} l_2\sigma^2 \right) \geq 1 - \frac{M_4}{n} - 4p^{-2}. \quad 90$$
Theorem A.2 (Concentration with light-tailed noise). Assume $x_i \sim N(0, \Sigma)$ with $p > c_0 n$ and $\varepsilon \sim N(0, \sigma^2)$. Define $\tilde{X} = (XX^T/p)^{-\frac{1}{2}}X$ and $\tilde{\varepsilon} = (XX^T/p)^{-\frac{1}{2}}\varepsilon$. Now for any $C > 0$, if $p / \log p \geq 16c_0/(\sqrt{c_0} - 1)^2$, then for any $t > 0$,

$$
\mathbb{P}\left( \frac{1}{n} \norm{\tilde{X} \tilde{\varepsilon}}_x \leq C_3 \frac{\sigma t}{\sqrt{n}} \right) \geq 1 - 2pe^{-t^2/2} - 2p^{-2} - 2pe^{-Cn},
$$

and

$$
\mathbb{P}\left( \frac{1}{2c^*v_2} \leq \frac{1}{n} \norm{\varepsilon}_2 \leq \frac{5}{2c^*v_1} \right) \geq 1 - 2p^{-2} - 2e^{-n/4}.
$$

where $C_3 = \sqrt{\frac{c_2c^*v_1}{c^*}}$ and $C, c_1, c_2, c_*, c^*$ are the same constants defined in Theorem 2.4 and $v_1$ is defined in Proposition 4.

Proof of Theorem A.2. Since $X \sim N(0, \Sigma)$, $X$ can be represented as $X = Z\Sigma^{1/2}$ where $Z$ is an $n \times p$ random matrix with i.i.d Gaussian entries. Thus, we have

$$
e_i^T \tilde{X} \tilde{\varepsilon} = e_i^T X^T (XX^T/p)^{-\frac{1}{2}} \varepsilon \sim N(0, \sigma^2 e_i^T X^T (XX^T/p)^{-2} X e_i).
$$

Defining $H = X^T (XX^T)^{-1/2}$ and using the results in the proof of Theorem 2.4, we have

$$
\max_i e_i^T X^T (XX^T/p)^{-2} X e_i \leq \max_i pe_i^T H H^T e_i \lambda_{\min}^{-1}(XX^T/p) 
\leq \frac{c_2c^*}{c^*} n \lambda_{\min}^{-1}(Z \Sigma Z^T/p) \leq \frac{c_2c^*}{c^*} n \lambda_{\min}^{-1}(ZZ^T/p) 
\leq \frac{c_2c^*}{c^*_2 v_1} n,
$$

with probability at least $1 - 2p^{-2} - 2pe^{-Cn}$. Therefore, using tail bound on the Gaussian variable, we have for any $t > 0$

$$
\mathbb{P}\left( \frac{1}{n} \norm{\tilde{X} \tilde{\varepsilon}}_x \leq \sqrt{\frac{c_2c^*}{c^*_2 v_1} \sigma t \sqrt{n}} \right) \geq 1 - 2pe^{-t^2/2} - 2p^{-2} - 2pe^{-Cn}.
$$
Now for $\frac{1}{n}\|\tilde{\varepsilon}\|_2^2$ we have that

$$\frac{1}{n}\|\tilde{\varepsilon}\|_2^2 = \frac{1}{n}\varepsilon^T(XX^T/p)^{-1}\varepsilon \in \left[\frac{1}{n} \frac{1}{c^*v_2} \|\varepsilon\|_2^2, \frac{1}{n} \frac{1}{c^*v_1} \|\varepsilon\|_2^2\right],$$

with probability at least $1 - 2p^{-2}$. Now using the bound on the $X^2(n)$ variable we have

$$\mathbb{P}\left(\frac{1}{c^*v_2} \leq \frac{1}{n}\|\varepsilon\|_2^2 \leq \frac{5}{2c_*v_1}\right) \geq 1 - 2p^{-2} - 2e^{-n/4}.$$

\[\square\]

### A.3 Proof of Lemma 3

We restate the Lemma 3 here.

**Lemma 8.** Assume $X \sim EN(L, \Sigma)$. If $p > c_0n$ for some $c_0 > 1$, then have for any $\gamma_1, \gamma_2 \subseteq [p]$ with $\gamma_1 \cap \gamma_2 = \emptyset$ and any $t > 0$, we have

$$\mathbb{P}\left(\frac{1}{n}\|\tilde{X}_{\gamma_1}^T \tilde{X}_{\gamma_2}\|_{\beta_{\gamma_1}\beta_{\gamma_2}} \leq \frac{c_4c^*\|\beta_{\gamma_2}\|_2}{c_*} \frac{t}{\sqrt{n}}\right) \geq 1 - 2|\gamma_1|e^{-t^2/2} - 5|\gamma_1|e^{-Cn},$$

**Proof.** To prove the second part of Lemma 3, we first define $H = X^T(XX^T)^{-\frac{1}{2}}$. When $X \sim EN(L, \Sigma)$, $H$ follows the $MACG(\Sigma)$ distribution as indicated in Lemma 3 in Wang et al. (2015b) and Theorem 1 in Wang and Leng (2015).

Define $k = |\gamma_1|$. Without loss of generosity, we assume $\gamma_1 = \{1, 2, 3, \cdots, k\}$ as we can always reorder the features so that features in $\gamma_1$ are placed in front of their complement. Then for any $p$-variate vector $v$ with $v_{\gamma_2} = 0$, we can always identify a $(p-k) \times (p-k)$ orthogonal matrix $T'$ such that $T'v_{\gamma_1} = v_{\gamma_2}\|\beta'_{\gamma_2}\|_1$ where $\beta'_{\gamma_2}$ is a $(p-k) \times 1$ unit vector with the first coordinate being 1. Now we define a new orthogonal matrix $T$ as

$$T = \begin{pmatrix} I_k & 0 \\ 0 & T' \end{pmatrix}$$
and we have

$$Tv = \left( I_k \ 0 \right) \left( \begin{array}{c} 0 \\ v_{(\gamma_1)_e} \end{array} \right) = \left( \begin{array}{c} 0 \\ v_{\gamma_2} \end{array} \right) = \|v_{\gamma_2}\|_{2e_{k+1}}.$$

and for any $i \in \gamma_1$ we have

$$e_i^T H H^T v = e_i^T T H H^T T^T T v = e_i^T T^T H H^T T^T e_{k+1} = \|v_{\gamma_2}\|_{2e_i^T \tilde{H} \tilde{H}^T e_{k+1}}.$$

Since $H$ follows $MACG(\Sigma)$, $\tilde{H} = T^T H$ follows $MACG(T^T \Sigma T)$ for any fixed $T$. Therefore, we can use Theorem 2.4 again to obtain that

$$\mathbb{P}\left(|e_i^T X^T (XX^T)^{-1} X v| \geq \frac{\|v_{\gamma_2}\|_{2c_4 c^* t} \sqrt{n}}{c_*} \right)$$

$$= \mathbb{P}\left(|v_{\gamma_2}\|_{2c_4 c^* t} \sqrt{n} \geq \frac{\|v_{\gamma_2}\|_{2c_4 c^* t} \sqrt{n}}{c_*} \right)$$

$$= \mathbb{P}\left(|v_{\gamma_2}\|_{2c_4 c^* t} \geq \frac{\|v_{\gamma_2}\|_{2c_4 c^* t} \sqrt{n}}{c_*} \right)$$

$$\leq 5e^{-Cn} + 2e^{-t^2/2}.$$

Applying the above result to $v_{\gamma_2} = \beta_{\gamma_2}^*$ we have

$$\frac{1}{n} \|X_{\gamma_1} \beta_{\gamma_2} \|_x = \max_{i \in \gamma_1} \frac{1}{n} |e_i^T X^T X v| = \frac{p}{n} |e_i^T X^T (XX^T)^{-1} X v| \leq \frac{c_4 c^* t \\|\beta_{\gamma_2}\|_2}{c_*} \sqrt{n},$$

with probability at least $1 - 5|\gamma_1|e^{-Cn} - 2|\gamma_1|e^{-t^2/2}$. \qed
Appendix B

Theory on the regularized regression

In Appendix B, we prove the consistency results in Section 2.2 for the regularized regression under deterministic conditions. In particular, we prove Theorem 2.1 and 2.2.

Before proceeding to the main proof, we first cite the following results from Fan and Lv (2011).

Lemma 9 (Local optimizer). The solution \( \hat{\beta} \) is a strict local minimizer of the penalized loss function \( l_n(\beta) \) if

1. \( \frac{1}{n} x_k^T (Y - X \hat{\beta}) = c \lambda_n \tilde{\rho}(\hat{\beta}_k), \ \forall \hat{\beta}_k \neq 0 \)

2. \( \frac{1}{cn \lambda_n} |x_k^T (Y - X \hat{\beta})| \leq 1, \ \forall \hat{\beta}_k = 0 \)

3. \( \lambda_{\min}(X_J^T X_J/n) \geq c \lambda_n \kappa_0(\rho; \hat{\beta}_J), \ J = \{k : \hat{\beta}_k \neq 0\}, \) where \( \kappa(\rho; v) \) is defined as

\[
\kappa_0(\rho; v) = \lim_{\epsilon \to 0} \max_{i \in [p]} \sup_{|v_i| - \epsilon < t_1 < t_2 < |v_i| + \epsilon} -\frac{\rho'(t_2) - \rho'(t_1)}{t_2 - t_1} \leq \sup_{0 < t_1 < t_2} -\frac{\rho'(t_2) - \rho'(t_1)}{t_2 - t_1} = \kappa_0(\rho)
\]

The value \( \kappa_0(\rho) \) can be computed easily for usual penalties. For example, for lasso, SCAD and MCP, we have \( \kappa_0(\rho) \) equals to 0, \((a - 1)^{-1}\) and \(a^{-1}\) respectively.
In addition, we need the following result to ensure the local sparse optimizer is also a global optimizer.

**Lemma 10** (Restricted global optimality). Assume there exists some constant $s$ such that $s < n/2$ and the design matrix $X$ satisfies that

$$\lambda_{\text{min}}\left(\frac{1}{n}X_T^TX\right) \geq \lambda_n\kappa_0(\rho)$$

for any $n \times 2s$ sub-matrices $X_J$ with $J \subseteq \{1, 2, \cdots, p\}$ and $|J| = 2s$, then the solution to $l_n(\beta)$ is a global minimizer among all $s$-sparse solutions.

Lemma 9 and 10 together ensure that there exists a unique local sparse optimizer of the loss function $l_n(\beta)$, allowing one to capture the behavior of the solution to the optimization problem. Therefore, we have the following result.

**Lemma 11.** Assume the penalty specification. Define $w = Y - X\beta*$, $\gamma_* = \{i : |\beta_i| \geq A\lambda_n\}$ and $s_* = |\gamma_*| \leq s_0$. We assume

1. $\max_{i \in [p]} M_1 \leq \frac{1}{n}\|x_i\|^2_2 \leq M_2$.
2. $\max_{i \neq j \in [p]} \frac{1}{n}|x_i^T x_j| \leq \tau_1\lambda_n$.
3. $\max_{i \in [p]} \frac{1}{n}|x_i^T w| \leq \tau_2\lambda_n$.
4. $\max_{i \notin \gamma_*} |\beta_{*i}| \leq B\lambda_n$ and $\sum_{i \notin \gamma_*} |\beta_{*i}| \leq R$.

If the following conditions hold

$$c \geq 3\tau_1 R + 3\tau_2 + 2BM_2, \quad A > 4(\tau_2 + \tau_1 R + c)/M_1 \quad \text{and} \quad \lambda_n \leq \min\left\{\frac{3M_1}{4c}, \frac{M_1}{4\tau_1 s_0}, \frac{M_1}{2\kappa_0(\rho)}\right\},$$

then there exists a unique sparse solution to the optimization problem $\hat{\beta}$ with at most $s_0$ nonzero coordinates satisfying that

$$\|\hat{\beta}_{\gamma_*} - \beta_{*\gamma_*}\|_\infty \leq A\lambda_n/2 \quad \text{and} \quad \hat{\beta}_{(\gamma_*)^c} = 0.$$
Proof of Lemma 11. Follow the proof in Fan and Lv (2011), we divide our proof in two steps. Define $\gamma_*$ to be the set containing all strong signals $\gamma_* = \{ i : |\beta_i| \geq A\lambda_n \}$ and $s_* = |\gamma_*|$

**Step 1.** In step 1, we show that there exists a solution $\hat{\beta}$ satisfying (1) and

$$
\| \hat{\beta}_{\gamma_*} - \beta_{\gamma_*} \| \leq A\lambda_n/2 \quad \text{and} \quad \hat{\beta}_{(\gamma_*)^c} = 0.
$$

Define a hypercube $N$ as $N = \{ v : \| v - \beta_{\gamma_*} \| \leq A\lambda_n/2 \}$ and we have $\min_{i \in [s_*]} |v_i| \geq A\lambda_n/2$ and $\text{sign}(v) = \text{sign}(\beta_{\gamma_*})$. Define $\eta(v) = c\lambda_n \tilde{\rho}(v)$, then we have

$$
\max_{v \in N} \| \eta(v) \| \leq c\lambda_n \rho' \left( A\lambda_n/2 \right) \leq c\lambda_n.
$$

Thus, for any $v \in N$ we have

$$
\max_{k \in \gamma_*} \left| \frac{1}{n} x_k^T Y - \frac{1}{n} x_k^T X_{\gamma_*} \beta_{\gamma_*} - \eta(v_k) \right| = \max_{k \in \gamma_*} \left| \frac{1}{n} x_k^T w + \frac{1}{n} x_k^T X_{(\gamma_*)^c} \beta_{(\gamma_*)^c} - \eta(v_k) \right|
\leq \tau_2 \lambda_n + R\tau_1 \lambda_n + c\rho'(A\lambda_n/2)\lambda_n = (\tau_2 + \tau_1 R + c)\lambda_n.
$$

Now we define a function $\Psi(v)$ as

$$
\Psi(v) = \frac{1}{n} X_{\gamma_*}^T X_{\gamma_*} v - \frac{1}{n} X_{\gamma_*}^T X_{\gamma_*} \beta_{\gamma_*} - \left( \frac{1}{n} X_{\gamma_*}^T Y - \frac{1}{n} X_{\gamma_*}^T X_{\gamma_*} \beta_{\gamma_*} - \eta(v_k) \right).
$$

Notice that $\Psi(v) = 0$ implies that

$$
\frac{1}{n} X_{\gamma_*}^T (Y - X_{\gamma_*} v) = \eta(v) = \lambda_n \tilde{\rho}(v),
$$

i.e., there exists a vector $v \in N$ such that $\hat{\beta}_{\gamma_*} = v$ and $\hat{\beta}_{(\gamma_*)^c} = 0$ will be a solution to (1). Therefore, we just need to verify that $\Psi(v) = 0$ has solution incide $N$. Taking $v_1 = \beta_{\gamma_*} + A\lambda_n/2$ and $v_2 = \beta_{\gamma_*} - A\lambda_n/2$, we have for any $k \in \gamma_*$

$$
e_k^T \Psi(v_1) \geq \frac{1}{n} x_k^T X_{\gamma_*} (v - \beta_{\gamma_*}) - (\tau_2 + \tau_1 R + c)\lambda_n
\leq (M_1 - s_* \tau_1 \lambda_n) A\lambda_n/2 - (\tau_2 + \tau_1 R + c)\lambda_n
\leq \left( \frac{1}{4} M_1 A - \tau_2 - \tau_1 R - c \right) \lambda_n
$$
and

\[ e_k^T \Psi(v_2) \leq \frac{1}{n} x_k^T X_{\gamma_*}^T (v - \beta_{\gamma_*}) + (\tau_2 + \tau_1 R + c) \lambda_n \]

\[ \leq - \left( \frac{1}{4} M_1 A - \tau_2 - \tau_1 R - c \right) \lambda_n \]

Therefore, as long as \( A > 4(\tau_2 + \tau_1 R + c)/M_1 \), we have \( \Psi(v_1) > 0 \) and \( \Psi(v_2) < 0 \).

By the continuity of the vector-valued function \( \Psi(v) \), there exists \( v \in \mathcal{N} \) such that \( \Psi(v) = 0 \). Define \( \hat{\beta} \) as

\[ \hat{\beta}_{\gamma_*} = v \quad \text{and} \quad \hat{\beta}_{(\gamma_*)^c}, \]

then \( \hat{\beta} \) is a valid solution to (1) and satisfies that \( \|\hat{\beta}_{\gamma_*} - \beta_{\gamma_*}\|_\infty \leq A \lambda_n/2 \).

**Step 2.** In Step 2, we verify the solution we found in Step 1 also satisfies (2) and (3). A direct calculation shows that for \( k \in (\gamma_*)^c \) we have

\[ \frac{1}{cn \lambda_n} |x_k^T (Y - X \hat{\beta})| = \frac{1}{cn \lambda_n} \left| x_k^T X_{\gamma_*} (\beta_{\gamma_*} - \hat{\beta}_{\gamma_*}) + x_k^T X_{(\gamma_*)^c} \beta_{(\gamma_*)^c} + x_k^T \right| \]

To bound the first term, we make use of (1) so that

\[ \frac{1}{n} X_{\gamma_*}^T X_{\gamma_*} (\beta_{\gamma_*} - \hat{\beta}_{\gamma_*}) + \frac{1}{n} X_{\gamma_*}^T X_{(\gamma_*)^c} \beta_{(\gamma_*)^c} + \frac{1}{n} X_{\gamma_*}^T \bar{w} = c \lambda_n \bar{\rho}(\hat{\beta}_{\gamma_*}) \]
and thus we have
\[
\frac{1}{n} |x_k^T X_{\gamma*} (\beta_{\gamma*} - \hat{\beta}_{\gamma*})| = |c \lambda_n x_k^T X_{\gamma*} (X_{\gamma*}^T X_{\gamma*})^{-1} \tilde{p}(\hat{\beta}_{\gamma*}) - \frac{1}{n} x_k^T X_{\gamma*} (X_{\gamma*}^T X_{\gamma*})^{-1} X_{\gamma*}^T X_{\gamma*} c \beta_{\gamma*}| \\
- \frac{1}{n} |x_k^T X_{\gamma*} (X_{\gamma*}^T X_{\gamma*})^{-1} X_{\gamma*}^T w|
\]
\[
\leq |c \lambda_n x_k^T X_{\gamma*} (X_{\gamma*}^T X_{\gamma*})^{-1} \tilde{p}(\hat{\beta}_{\gamma*})| + \frac{1}{n} |x_k^T X_{\gamma*} (X_{\gamma*}^T X_{\gamma*})^{-1} X_{\gamma*}^T X_{\gamma*} c \beta_{\gamma*}| \\
+ \frac{1}{n} |x_k^T X_{\gamma*} (X_{\gamma*}^T X_{\gamma*})^{-1} X_{\gamma*}^T w|
\]
\[
\leq c \lambda_n \|x_k^T X_{\gamma*}\|_1 \|(X_{\gamma*}^T X_{\gamma*})^{-1}\|_\infty + \frac{1}{n} \|x_k^T X_{\gamma*}\|_1 \|(X_{\gamma*}^T X_{\gamma*})^{-1}\|_\infty \|X_{\gamma*}^T X_{\gamma*} c \beta_{\gamma*}\|_\infty \\
+ \frac{1}{n} \|x_k^T X_{\gamma*}\|_1 \|(X_{\gamma*}^T X_{\gamma*})^{-1}\|_\infty \|X_{\gamma*}^T w\|_\infty
\]
\[
\leq \frac{c \tau_1 s_* \lambda_n^2}{M_1 - \tau_1 s_* \lambda_n} + \frac{\gamma_1^2 s_* R \lambda_n^2}{M_1 - \tau_1 s_* \lambda_n} + \frac{\tau_1 \tau_2 s_* \lambda_n^2}{M_1 - \tau_1 s_* \lambda_n}
\]
\[
= \left(c + \tau_1 R + \tau_2\right) \frac{2 \tau_1 s_* \lambda_n^2}{M_1}
\]

For the second term, we notice that
\[
\frac{1}{n} |x_k^T X_{\gamma*} c \beta_{\gamma*}| \leq \frac{1}{n} \|x_k\|_2^2 |\beta_{\gamma*}| + \sum_{j \neq k} \frac{1}{n} |x_j^T x_k \beta_{\gamma*}| \leq (BM_2 + \tau_1 R) \lambda_n
\]

Together we have
\[
\frac{1}{cn \lambda_n} |x_k^T (Y - X \hat{\beta})| \leq \left(c + \tau_1 R + \tau_2\right) \frac{2 \tau_1 s_* \lambda_n}{c M_1} + \frac{BM_2 + \tau_1 R}{c} + \frac{\tau_2}{c}
\]
\[
\leq \frac{1}{2} + \frac{3 \tau_1 R + 3 \tau_2 + 2BM_2}{2c}.
\]

Therefore, as long as \( c \geq 3 \tau_1 R + 3 \tau_2 + 2BM_2 \), the solution defined in Step 1 satisfies (2). For (3), it is automatically satisfied from our assumption that \( \tau_1 s_* \lambda_n \leq M_1/4 \), since
\[
\lambda_{\text{min}}(X_{\gamma*}^T X_{\gamma*}/n) \geq \min_{i \in \gamma*} \frac{1}{n} \|x_i\|_2^2 - \sum_{j \neq i} \frac{1}{n} |x_j^T x_i| \geq M_1 - \tau_1 s_* \lambda_n \geq \frac{3}{4} M_1
\]
with the condition $\lambda_n \leq (3M_1)/(4c)$.

In Lemma 11, we make use of the following inequality
\[
\|X_{\gamma_\ast}^T X_{\gamma_\ast} \beta_{\gamma_\ast} \|_\infty \leq \|X_{\gamma_\ast}^T X_{\gamma_\ast} \beta_{\gamma_\ast} \|_1 \leq \tau_1 R \lambda_n
\]
to prove the final result, which brings $R$ into the signal strength in the conditions.

In some cases where $\|X_{\gamma_\ast}^T X_{\gamma_\ast} \beta_{\gamma_\ast} \|_\infty$ can be bounded in a better way, we might be able to remove the $\ell_1$ constraint on the weak signals. Thus, we present another Lemma for such case

**Lemma 12.** Assume the penalty specification. Define $w = Y - X \beta_\ast$, $\gamma_\ast = \{i : |\beta_i| \geq A \lambda_n\}$ and $s_\ast = |\gamma_\ast| < s_0$. We assume

1. $\max_{i \in [p]} M_1 \leq \frac{1}{n} \|x_i\|_2^2 \leq M_2$.
2. $\max_{i \neq j \in [p]} \frac{1}{n} |x_i^T x_j| \leq \tau_1 \lambda_n$.
3. $\max_{i \in [p]} \frac{1}{n} |x_i^T w| \leq \tau_2 \lambda_n$.
4. $\max_{i \neq \gamma_\ast} |\beta_{\gamma_\ast}| \leq B \lambda_n$ and $\|X_{\gamma_\ast}^T X_{\gamma_\ast} \beta_{\gamma_\ast} \|_\infty \leq \tau_3 \lambda_n$.

If the following conditions hold

\[
c \geq 3\tau_3 + 3\tau_2 + 2BM_2, \quad A > 4(\tau_2 + \tau_3 + c)/M_1 \quad \text{and} \quad \lambda_n \leq \min \left\{ \frac{3M_1}{4c}, \frac{M_1}{4\tau_1 s_0}, \frac{M_1}{2\kappa_0(\rho)} \right\},
\]

then there exists a unique sparse solution to the optimization problem $\hat{\beta}$ with at most $s_0$ nonzero coordinates satisfying that

\[
\|\hat{\beta}_{\gamma_\ast} - \beta_{\gamma_\ast}\|_\infty \leq A \lambda_n/2 \quad \text{and} \quad \hat{\beta}_{(\gamma_\ast)}^c = 0.
\]

To this end, we have proved that there exists a local optimizer that possesses a good feature selection property. The uniqueness of this sparse solution is guaranteed by Lemma 10. As long as $\lambda_n \leq \frac{M_1}{2s_0(\rho)}$, $\hat{\beta}$ will be the only solution that has at most $s_0$ nonzero coordinates, which completes the whole proof.
The proof of Lemma 12 follows exactly the same procedure as Lemma 11 with all the $\tau_1 R$ replaced by $\tau_3$.

The Theorem 2.1 and 2.2 are direct results of Lemma 11 and 12, i.e., we have the following.

**Theorem B.1.** Assume all the conditions in Lemma 11 with $\lambda_n^2 = \sigma^2 \log p \, n$ and the penalty conditions. If the following holds

$$c \geq 3\tau_1 R + 3\tau_2 + 2BM_2, \quad A > 4(\tau_2 + \tau_1 R + c)/M_1$$

and

$$n \geq \max \left\{ \frac{16c^2}{9M_1^2}, \frac{16\tau_1^2 s_0^2}{M_1^2}, \frac{4\kappa_0(\rho)}{M_1^2} \right\} \log p,$$

then there exists a unique sparse solution to the optimization problem $\hat{\beta}$ with at most $s_0$ nonzero coordinates satisfying that

$$\|\hat{\beta}_{\gamma^*} - \beta_{\gamma^*}\|_\infty \leq \frac{A\sigma}{2} \sqrt{\frac{\log p}{n}} \quad \text{and} \quad \hat{\beta}_{(\gamma^*)^c} = 0.$$

and

**Theorem B.2.** Assume all the conditions in Lemma 12 with $\lambda_n^2 = \sigma^2 \log p \, n$ and the penalty conditions. If the following holds

$$c \geq 3\tau_3 + 3\tau_2 + 2BM_2, \quad A > 4(\tau_2 + \tau_3 + c)/M_1$$

and

$$n \geq \max \left\{ \frac{16c^2}{9M_1^2}, \frac{16\tau_1^2 s_0^2}{M_1^2}, \frac{4\kappa_0(\rho)}{M_1^2} \right\} \log p,$$

then there exists a unique sparse solution to the optimization problem $\hat{\beta}$ with at most $s_0$ nonzero coordinates satisfying that

$$\|\hat{\beta}_{\gamma^*} - \beta_{\gamma^*}\|_\infty \leq \frac{A\sigma}{2} \sqrt{\frac{\log p}{n}} \quad \text{and} \quad \hat{\beta}_{(\gamma^*)^c} = 0.$$
Appendix C

Theory on the Bayesian variable selection

In Appendix C, we prove the consistency results in Section 2.3 for the Bayesian variable selection under deterministic conditions. In particular, we prove Theorem 2.3. This chapter consists of two parts. In the first part, we revise the consistency theory in Yang et al. (2015) to be deterministic to better fit under our proof framework. In the second part, we verify these deterministic conditions on under Condition 2.

C.1 Deterministic conditions for Bayesian variable selection

From the prior specification (Condition 3) we know that the prior mass is located on the model set $\mathcal{M}_{s_0} = \{ \gamma : |\gamma| \leq s_0 \}$. For any of the following statement, we will assume the models are within $\mathcal{M}_{s_0}$. Let $\Phi_A$ denote the projection matrix formed by submatrix $X_A$, i.e., $\Phi_A = X_A(X_A^TX_A)^{-1}X_A$. We need following conditions,

1. The error $w = Y - X\beta_*$ satisfies that For any two models $\gamma_1 \subset \gamma_2$ and for any $\gamma$

$$\frac{1}{n} w^T(\Phi_{\gamma_2} - \Phi_{\gamma_1})w \leq (|\gamma_2| - |\gamma_1|)L\lambda_n^2$$
and

$$\frac{w^T\Phi_{\gamma}w}{n|\gamma|} \leq q\lambda_n^2$$
2. The coefficients satisfy that for any $\gamma_2 = \gamma_1 \cup \gamma_*$ and any $\gamma$

$$
\frac{1}{n}\|(\Phi_{\gamma_2} - \Phi_{\gamma_1})X_{(\gamma_*)^c}\beta_{(\gamma_*)^c}\|^2_2 \leq (|\gamma_2| - |\gamma_1|)L_1\lambda_n^2, \quad \frac{1}{n}\|w^T(I-\Phi_{\gamma})X_{(\gamma_*)^c}\beta_{(\gamma_*)^c}\| \leq \sigma L_2\lambda_n
$$

and $\min_{i\in\gamma_*} |\beta_{\gamma_i}| \geq A\lambda_n$.

3. We need a quantification on the magnitude of the error $w$, the weak and the strong signals

$$
r'\sigma^2 \leq \frac{1}{n}\|w\|^2_2 \leq r\sigma^2, \quad \frac{1}{n}\|X_{(\gamma_*)^c}\beta_{(\gamma_*)^c}\|^2_2 \leq q_1\sigma^2 \quad \text{and} \quad \frac{1}{n}\|X_{\gamma_}\beta_{\gamma_}\|^2_2 \leq q_2g\sigma^2
$$

4. For all models $|\gamma| \leq s_0$, we have $\frac{1}{n}\lambda_{\min}(X_{\gamma_}^TX_{\gamma}) \geq \nu^2$.

The following lemma characterizes the posterior probability for models inside this class.

**Lemma 13.** Assume the four conditions. If $\lambda_n^2 \leq \min\left\{\frac{r\sigma^2}{4qs_0}, \frac{r^2\sigma^2}{64L_2^2}\right\}$, $\alpha > \frac{(2L+2L_1)n}{r\sigma^2\log p} \lambda_n^2 + 3 - \kappa, 12q_1 + 12r \leq q_2g^2$ and $\nu^2A_2\lambda_n^2 \geq 16(4q_1 + 7q_2 + 4r)\{\alpha + \kappa + 5\}\sigma^2\log p/n$, then the posterior probability of models $\gamma \neq \gamma_*$ satisfy

$$
\frac{\pi_n(\gamma|Y)}{\pi_n(\gamma_*|Y)} \leq \begin{cases} p^{-3|\gamma|\gamma_*|} & \text{if } \gamma_* \subset \gamma \\ p^{-3|\gamma|^{-2}} & \text{otherwise} \end{cases}
$$

**Proof of Lemma 13.** Case 1. $\gamma_* \subset \gamma$. Let $k = |\gamma| \leq s_0$ We follow Yun’s proof we obtain that

$$
1 - R_n^2 = \frac{\|(1 - \Phi_{\gamma})X_{(\gamma_*)^c}\beta_{(\gamma_*)^c} + w\|^2_2}{\|Y\|^2_2} \geq \frac{\|(1 - \Phi_{\gamma})w\|^2_2 - 2|w^T(I-\Phi_{\gamma})X_{(\gamma_*)^c}\beta_{(\gamma_*)^c}|}{\|Y\|^2_2} \geq \frac{\frac{1}{n}\|(I - \Phi_{\gamma})w\|^2_2 - 2\sigma L_2\lambda_n}{\|Y\|^2_2/n}.
$$
Using the fact that $\gamma \subset \gamma$, i.e., $\Phi_\gamma - \Phi_{\gamma_*}$ is also a projection matrix, we have

$$R^2_\gamma - R^2_{\gamma_*} \leq \frac{2}{n} \| (\Phi_\gamma - \Phi_{\gamma_*}) w \|^2 + 2kL_1 \lambda^2_n / \|Y\|^2 / n.$$ 

Now notice that we have

$$\frac{1}{n} \| (I - \Phi_\gamma) w \|^2 = \frac{1}{n} \| w \|^2 - \frac{w^T \Phi_\gamma w}{n} \geq r' \sigma^2 - qk \lambda^2_n \geq r' \sigma^2 - qs_0 \lambda^2_n$$

and

$$\frac{1}{n} \| w^T (\Phi_\gamma - \Phi_{\gamma_*}) w \|^2 \leq kL \lambda^2_n.$$ 

Using the assumption, we have $qs_0 \lambda^2_n \leq r' \sigma^2 / 4$ and $2 \sigma L_2 \lambda_n \leq r' \sigma^2 / 4$,

$$1 - R^2_\gamma \geq \frac{r' \sigma^2 - qs_0 \lambda^2_n - 2 \sigma L_2 \lambda_n}{\|Y\|^2 / n} \geq \frac{r' \sigma^2 / 2}{\|Y\|^2 / n}$$

and

$$R^2_\gamma - R^2_{\gamma_*} \leq \frac{2kL \lambda^2_n + 2kL_1 \lambda^2_n}{\|Y\|^2 / n} = \frac{(2kL + 2kL_1) \lambda^2_n}{\|Y\|^2 / n}.$$

As a result, assuming $\alpha > (2L + 2L_1) \frac{n \lambda^2_n}{r' \sigma^2 \log p} + 3 - \kappa$, we have

$$\frac{\pi_n(\gamma \mid Y)}{\pi_n(\gamma_* \mid Y)} = \frac{1}{p^{nk}(1 + g)^{k/2}} \left( 1 + \frac{(2kL + 2kL_1) \lambda^2_n}{r' \sigma^2 / 2} \right)^{\frac{n}{2}}$$

$$\leq \frac{1}{p^{nk}(1 + g)^{k/2}} \exp \left( \frac{2kL + 2kL_1}{r' \sigma^2} \frac{n \lambda^2_n}{\log p} \right)$$

$$\leq \frac{1}{p^{nk}(1 + g)^{k/2}} \exp \left( \frac{(2kL + 2kL_1)n \lambda^2_n}{r' \sigma^2 \log p} \right)$$

$$= p^{k(2L + 2L_1) \frac{n \lambda^2_n}{r' \sigma^2 \log p} - k(\alpha + \kappa)} \leq p^{-3k} = p^{-3(|\gamma \setminus \gamma|)}.$$ 

**Case 2.** $\gamma_* \notin \gamma$. Follow Yun’s paper, we define $\tilde{\gamma} = \gamma_* \cup \gamma$ and $\tilde{k} = |\tilde{\gamma} \setminus \gamma| = |\gamma_* \setminus \gamma| \geq 1$. We also define $l = |\gamma|$ and thus $|\tilde{\gamma} \setminus \gamma_*| = l + \tilde{k} - s_*$. With the same
Thus, as long as
\[ 1 - R_\gamma^2 - (1 - R_\gamma^2) \geq \left( \frac{\| (\Phi_\gamma - \Phi_\gamma) X_{\gamma_\beta, \beta_{\gamma_\beta}} \|_2}{\| Y \|_2} - \frac{\| (\Phi_\gamma - \Phi_\gamma) X_{\gamma_\beta, \beta_{\gamma_\beta}} \|_2}{\| Y \|_2} - \frac{\| (\Phi_\gamma - \Phi_\gamma) w \|_2}{\| Y \|_2} \right)^2 \]

From the proof in Case 1 we know that
\[ \frac{1}{n} \| (\Phi_\gamma - \Phi_\gamma) w \|_2^2 \leq \tilde{k} L \lambda_n^2 \quad \text{and} \quad \frac{1}{n} \| (\Phi_\gamma - \Phi_\gamma) X_{\gamma_\beta, \beta_{\gamma_\beta}} \|_2^2 \leq \tilde{k} L_1 \lambda_n^2 \]

and using the Condition 4, we have
\[ \frac{1}{n} \| (\Phi_\gamma - \Phi_\gamma) X_{\gamma_\beta, \beta_{\gamma_\beta}} \|_2 = \frac{1}{n} \| (I - \Phi_\gamma) X_{\gamma_\beta, \beta_{\gamma_\beta}} \|_2 \geq \nu^2 \| \beta_{\gamma_\beta} \|_2 \geq \nu^2 \tilde{k} A^2 \lambda_n^2. \]

Thus, as long as \( \nu^2 A^2 / 2 \geq L + L_1 \), we have
\[ \frac{1}{n} \| (\Phi_\gamma - \Phi_\gamma) w \|_2^2 + \frac{1}{n} \| (\Phi_\gamma - \Phi_\gamma) X_{\gamma_\beta, \beta_{\gamma_\beta}} \|_2^2 \leq \frac{1}{2n} \| (\Phi_\gamma - \Phi_\gamma) X_{\gamma_\beta, \beta_{\gamma_\beta}} \|_2^2 \quad (C.1) \]

and therefore
\[ 1 - R_\gamma^2 - (1 - R_\gamma^2) \geq \frac{\| (\Phi_\gamma - \Phi_\gamma) X_{\gamma_\beta, \beta_{\gamma_\beta}} \|_2^2}{4 \| Y \|_2^2} = \frac{\| (I - \Phi_\gamma) X_{\gamma_\beta, \beta_{\gamma_\beta}} \|_2^2}{4 \| Y \|_2^2} \]

On the other hand, for \( 1 - R_\gamma^2 \) we have
\[ 1 - R_\gamma = \frac{Y^T (I - \Phi_\gamma) Y}{\| Y \|_2^2} \]
\[ = \left( \frac{\| (I - \Phi_\gamma) X_{\gamma_\beta, \beta_{\gamma_\beta}} \|_2 + \| (I - \Phi_\gamma) X_{\gamma_\beta, \beta_{\gamma_\beta}} \|_2 + \| (I - \Phi_\gamma) w \|_2}{\| Y \|_2} \right)^2 \]
\[ \leq \frac{2 \| (I - \Phi_\gamma) X_{\gamma_\beta, \beta_{\gamma_\beta}} \|_2^2 + 4 \| I - \Phi_\gamma \|_2 \| X_{\gamma_\beta, \beta_{\gamma_\beta}} \|_2^2 + 4 \| (I - \Phi_\gamma) w \|_2^2}{\| Y \|_2^2} \]
\[ \leq \frac{2 \| (I - \Phi_\gamma) X_{\gamma_\beta, \beta_{\gamma_\beta}} \|_2^2 + (4q_1 + 4r)\sigma^2}{\| Y \|_2^2} \]

and \( Y \) satisfies that
\[ \| Y \|_2^2 \leq (\| X_{\gamma_\beta, \beta_{\gamma_\beta}} \|_2^2 + \| X_{\gamma_\beta, \beta_{\gamma_\beta}} \|_2^2 + \| w \|_2^2 \leq 3(2q_2 g + 4q_1 + 4r)\sigma^2 \leq 7q_2 g \sigma^2, \]
when \( g = p^{2\alpha} \) is sufficiently large. Based on the above result, we can compute the ratio between the posterior probability

\[
\frac{\pi_n(\gamma|Y)}{\pi_N(\gamma|Y)} = p^{\alpha k}(1 + g)^{\frac{k}{2}} \left( 1 - \frac{1 - R_\gamma^2 - (1 - R_{\gamma*}^2)}{g^{-1} + 1 - R_{\gamma*}^2} \right)^{\frac{n}{2}}
\]

\[
\leq p^{\alpha k}(1 + g)^{\frac{k}{2}} \left( 1 - \frac{\| (1 - \Phi_\gamma) X_{\gamma*} \beta_{\gamma*} \|^2}{4g^{-1} Y^n + 4(4q_1 + 4r)n\sigma^2 + 8\| (1 - \Phi_\gamma) X_{\gamma*} \beta_{\gamma*} \|^2} \right)^{\frac{n}{2}}
\]

\[
\leq p^{\alpha k}(1 + g)^{\frac{k}{2}} \left( 1 - \frac{\| (1 - \Phi_\gamma) X_{\gamma*} \beta_{\gamma*} \|^2}{4(4q_1 + 7q_2 + 4r)n\sigma^2 + 8\| (1 - \Phi_\gamma) X_{\gamma*} \beta_{\gamma*} \|^2} \right)^{\frac{n}{2}}
\]

\[
\leq p^{\alpha k}(1 + g)^{\frac{k}{2}} \left( 1 - \min \left( \frac{1}{16}, \frac{\nu^2 k A^2 \lambda_0^2}{8(4q_1 + 7q_2 + 4r)\sigma^2} \right) \right)^{\frac{n}{2}}
\]

According to our assumption that \( \nu^2 A^2 \lambda_0^2 \geq 16(4q_1 + 7q_2 + 4r)(\alpha + \kappa + 5)\sigma^2 \log p/n, \) thus

\[
\frac{\pi_n(\gamma|Y)}{\pi_N(\gamma|Y)} \leq p^{\alpha k}(1 + g)^{\frac{k}{2}} \left( 1 - 2(\alpha + \kappa + 5)\frac{k\log p}{n} \right)^{\frac{n}{2}} \leq p^{\alpha k}p^{(\alpha + \kappa + 5)\frac{k}{n}} = p^{-5\frac{k}{n}}.
\]

Now using result from Case 1, we know that

\[
\frac{\pi_n(\tilde{\gamma}|Y)}{\pi_N(\gamma|Y)} \leq p^{-3|\gamma \setminus \gamma*|} = p^{-3|\gamma \setminus \gamma*|}.
\]

Combining these two results, we get

\[
\frac{\pi_n(\gamma|Y)}{\pi_N(\gamma|Y)} \leq p^{-3|\gamma \setminus \gamma*| - 5\frac{k}{n}} \leq p^{-3|\gamma \setminus \gamma*| - 3|\gamma \setminus \gamma*| - 2} = p^{-3|\gamma| - 2}
\]

\[
\square
\]

Under Lemma 13, it is easy to show the posterior concentration on the true model. Models in \( \mathcal{M}_{s_0} \) can be classified into the two cases. For models in Case 1 satisfying \( |\gamma \setminus \gamma*| = k \), the total number will at most \( p^k \). On the other hand, for models in Case 2 satisfying \( |\gamma| = l \), the total number will also be at most \( p^l \). As a result, we have

\[
\frac{\pi_n(\mathcal{M}_{s_0}|Y)}{\pi_n(\gamma|Y)} \leq \sum_{k=1}^{s_0} p^k p^{-3k} + \sum_{l=0}^{s_0} p^l p^{-3l-2} \leq 3p^{-2}.
\]

105
Therefore, we have the following result

**Corollary C.1.** *Under the conditions of Lemma 13, we have that*

\[
\pi_n(\gamma_* | Y) \geq 1 - 3p^{-2}.
\]

### C.2 Relating to Condition 2

The conditions on Lemma 13 is hard to verify in many scenarios. To ease the proof for our main results, we simplify these conditions in this section that guarantee the validity of the original ones. The set of the new conditions are proposed in a way similar to their frequentist counterparts. We assume the coefficients \( \beta_* \) can be divided into the strong and the weak signals, where the strong signals are lower bounded by some threshold

\[
\gamma_* = i : |\beta_{*i}| \geq A\lambda_n \quad \text{and} \quad |\gamma_*| = s_* \leq s_0
\]

Now we have the following result

**Lemma 14.** *Let \( w = Y - X\beta_* \) and define \( \gamma_* \) and \( R \) as above. We assume the following conditions*

1. \( \max_{i \in [p]} M_1 = \frac{1}{n} \| x_i \|_2^2 \leq M_2. \)
2. \( \max_{i \neq j \in [p]} \frac{1}{n} |x_i^T x_j| \leq \tau_1 \lambda_n. \)
3. \( \max_{i \in [p]} \frac{1}{n} |x_i^T w| \leq \tau_2 \lambda_n. \)
4. *Same as the Condition 3 in Lemma 13 and \( \sum_{i \neq \gamma_*} |\beta_{*i}| \leq R. \)*

*If \( s_0 \tau_1 \lambda_n \leq M_1/2, \) then the Conditions in Lemma 13 are satisfied with*

\[
L = \frac{8\tau_2^2}{M_1}, \quad q = \frac{2\tau_2^2}{M_1}, \quad L_1 = \frac{13R^2 \tau_1^2}{M_1}, \quad L_2 = \frac{3\tau_2 R}{\sigma} \quad \text{and} \quad \nu^2 = \frac{M_1}{2}.
\]
Proof of Lemma 14. We verify Condition 1, 2, 4 based on the new set of conditions.

**Condition 1.** Notice that \( \frac{1}{n}w^T(\Phi_\gamma_2 - \Phi_\gamma_1)w \) can be decomposed as sums of nested projections, i.e., we can find a series of model \( \tilde{\gamma}_1, \tilde{\gamma}_2, \cdots, \tilde{\gamma}_k \) such that \( \gamma_1 \subset \tilde{\gamma}_1 \subset \tilde{\gamma}_2 \subset \cdots \subset \tilde{\gamma}_k \subset \gamma_2 \) and \( |\tilde{\gamma}_{i+1}| - |\tilde{\gamma}_i| = 1 \), and the original quantity can be expressed as

\[
\frac{1}{n}w^T(\Phi_\gamma_2 - \Phi_\gamma_1)w = \frac{1}{n}w^T(\Phi_\gamma_1 - \Phi_\gamma_1)w + \frac{1}{n}w^T(\Phi_\gamma_2 - \Phi_\gamma_1)w + \cdots + \frac{1}{n}w^T(\Phi_\gamma_2 - \Phi_\gamma_k)w.
\]

Therefore, we just need to prove the special case where \( |\gamma_2| - |\gamma_1| = 1 \) and the general result follow from the induction. Without loss of generosity, we assume \( \gamma_2 \setminus \gamma_1 = \{j\} \).

We make use of Lemma 6 in Yun et al. and have

\[
w^T(\Phi_\gamma_2 - \Phi_\gamma_1)w = \frac{|w^T(I - \Phi_\gamma_1)x_j|^2}{x_j^T(I - \Phi_\gamma_1)x_j}.
\]

It follows from our assumption that

\[|w^T x_j| \leq n\tau_2 \lambda_n \quad \text{and} \quad x_j^T x_j \geq M_1 n.\]

For \( w^T \Phi_\gamma_1 x_j \), notice that \( s_0 \tau_1 \lambda_n \leq \frac{M_1}{2} \), we have

\[
|w^T \Phi_\gamma_1 x_j| = |w^T X_{\gamma_1}(X_{\gamma_1}^T X_{\gamma_1})^{-1} X_{\gamma_1}^T x_j| \leq \|w^T X_{\gamma_1}\|_\infty \|(X_{\gamma_1}^T X_{\gamma_1})^{-1} X_{\gamma_1}^T x_j\|_1
\]

\[\leq n\tau_2 \lambda_n \|(X_{\gamma_1}^T X_{\gamma_1})^{-1}\|_\infty \|X_{\gamma_1}^T x_j\|_1
\]

\[\leq n\tau_2 \lambda_n (M_1 - s_0 \tau_1 \lambda_n)^{-1} s_0 \tau_1 \lambda_n
\]

\[\leq \frac{2s_0 \tau_1 \tau_2 \lambda_n^2}{M_1} \leq n\tau_2 \lambda_n,
\]

where we use the fact that for any matrix \( Q \), \( \|Q^{-1}\|_\infty \leq 1/\min_i (Q_{ii} - \sum_{j \neq i} Q_{ij}) \).

Follow a similar argument, we have for \( x_j^T \Phi_\gamma_1 x_j \)

\[
x_j^T \Phi_\gamma_1 x_j = x_j^T X_{\gamma_1}(X_{\gamma_1}^T X_{\gamma_1})^{-1} X_{\gamma_1}^T x_j \leq \|x_j^T X_{\gamma_1}\|_\infty \|(X_{\gamma_1}^T X_{\gamma_1})^{-1}\|_\infty \|X_{\gamma_1}^T x_j\|_1
\]

\[\leq n\tau_1 \lambda_n (M_1 - s_0 \tau_1 \lambda_n)^{-1} s_0 \tau_1 \lambda_n \leq \frac{2s_0 \tau_1 \lambda_n^2}{M_1} \leq M_1 n/2,
\]
whenever \( s_0 \geq 1 \). Consequently, we have

\[
\frac{1}{n} w^T (\Phi_{\gamma_2} - \Phi_{\gamma_1}) w = \frac{1}{n} \frac{x_j^T (I - \Phi_{\gamma_1}) x_j}{x_j^T (I - \Phi_{\gamma_1}) x_j} = \frac{1}{n} \frac{2|x_j^T x_j|^2 + 2|w^T (I - \Phi_{\gamma_1}) x_j|^2}{x_j^T (I - \Phi_{\gamma_1}) x_j} \\
\leq 4\tau^2 \lambda_n^2 \leq \frac{8\tau^2}{M_1/2} = \frac{8\tau^2}{M_1} \lambda_n^2
\]

Taking \( L = 8\tau^2/M_1 \) concludes the first part of Condition 1.

The second part of Condition 1 follows almost identical argument

\[
w^T \Phi_{\gamma} w = w^T X_{\gamma_1} (X_{\gamma_1}^T X_{\gamma_1})^{-1} X_{\gamma_1}^T w \leq \| w^T X_{\gamma} \|_2 \| (X_{\gamma}^T X_{\gamma})^{-1} \|_2 \| X_{\gamma} w \|_1 \\
\leq n\tau^2 (M_1 - s_0 \tau \lambda_n)^{-1} |\gamma| \lambda^2_n \leq \frac{2n \tau^2 |\gamma| \lambda^2_n}{M_1}
\]

and this concludes our proof for the second part with \( q = 2\tau^2/(M_1) \).

**Condition 2.** Verification of Condition 2 is almost identical to Condition 1 using our assumptions. For the first part, with the same argument, we only need to consider the case where \( \gamma_2 = \gamma_1 \cup \{j\} \) and we have

\[
\beta_{\ast(\gamma_2)}^T X_{\gamma_2}^T (\Phi_{\gamma_2} - \Phi_{\gamma_1}) X_{\gamma_2} \beta_{\ast(\gamma_2)} = \frac{|\beta_{\ast(\gamma_2)}^T X_{\gamma_2}^T (I - \Phi_{\gamma_1}) x_j|^2}{x_j^T (I - \Phi_{\gamma_1}) x_j}
\]

The denominator can be bounded as in Condition 1. For the numerator, noticing that \( j \notin (\gamma_2)^c \) we have

\[
|\beta_{\ast(\gamma_2)}^T X_{\gamma_2}^T x_j| \leq \| \beta_{\ast(\gamma_2)}^T \|_1 \| X_{\gamma_2}^T x_j \|_2 \leq nR \tau \lambda_n
\]

To quantify the second part, we define \( \gamma_{11} = \gamma_1 \cap (\gamma_2)^c \) and \( \gamma_{12} = (\gamma_1)^c \cap (\gamma_2)^c \), the
we have
\[ |\beta^T_{\gamma_1} X^T_{\gamma_1} \Phi_{\gamma_1} x_j| = |(\beta^T_{\gamma_1} X^T_{\gamma_1} + \beta^T_{\gamma_1} X^T_{\gamma_1}) \Phi_{\gamma_1} x_j| \]
\[ \leq |\beta^T_{\gamma_1} X^T_{\gamma_1} \Phi_{\gamma_1} x_j| + |\beta^T_{\gamma_1} X^T_{\gamma_1} \Phi_{\gamma_1} x_j| \]
\[ = |\beta^T_{\gamma_1} X^T_{\gamma_1} \Phi_{\gamma_1} x_j| + |\beta^T_{\gamma_1} X^T_{\gamma_1} \Phi_{\gamma_1} x_j| \]
\[ \leq n R \tau_1 \lambda_n + |\beta^T_{\gamma_1} X^T_{\gamma_1} (X^T_{\gamma_1} X^T_{\gamma_1})^{-1} X^T_{\gamma_1} X^T_{\gamma_1} \Phi_{\gamma_1} x_j| \]
\[ \leq n R \tau_1 \lambda_n + \|\beta^T_{\gamma_1} \|\|X^T_{\gamma_1} X^T_{\gamma_1} \Phi_{\gamma_1} x_j| \][X^T_{\gamma_1} X^T_{\gamma_1}^{-1} \|X^T_{\gamma_1} x_j| \[ \leq n R \tau_1 \lambda_n + n R s_0 \tau_1 \lambda_n (M_1 - s_0 \tau_1 \lambda_n)^{-1} \lambda_n / 2 \]
\[ \leq n R \tau_1 \lambda_n + \frac{n R s_0 \tau_1 \lambda_n^2}{M_1} \leq \frac{3}{2} n R \tau_1 \lambda_n \]

Consequently, we have
\[
\frac{1}{n} \|(\Phi_{\gamma_2} - \Phi_{\gamma_1}) X_{(\gamma_2)} \Phi_{(\gamma_2)}\|_2^2 \leq \frac{1}{n} \frac{2|\beta^T_{\gamma_1} X^T_{\gamma_1} \Phi_{\gamma_1} x_j|^2 + 2|\beta^T_{\gamma_1} X^T_{\gamma_1} \Phi_{\gamma_1} x_j|^2}{x_j^T (I - \Phi_{\gamma_1}) x_j} \]
\[ \leq \frac{13/2 R^2 \tau^2_1 \lambda_n^2}{M_1/2} = \frac{13 R^2 \tau_1^2}{M_1} \lambda_n^2 \]

and this concludes the first part with \( L_1 = 13 R^2 \tau_1^2 / M_1 \).

With a similar definition for \( \gamma_{11} \) and \( \gamma_{12} \) we have
\[ |w^T \Phi_{\gamma_1} X_{(\gamma_2)} \Phi_{(\gamma_2)}| = |w^T \Phi_{\gamma_1} X_{\gamma_{11}} \beta_{\gamma_{11}} + w^T \Phi_{\gamma_1} X_{\gamma_{12}} \beta_{\gamma_{12}}| \]
\[ \leq |w^T X_{\gamma_{11}} \beta_{\gamma_{11}}| + |w^T \Phi_{\gamma_1} X_{\gamma_{12}} \beta_{\gamma_{12}}| \]
\[ \leq \|w^T X_{\gamma_{11}} \| \|\beta_{\gamma_{11}}\|_1 + \|w^T X_{\gamma_{12}} \| \|X^T_{\gamma_{11}} X^T_{\gamma_{12}} \| \|\beta_{\gamma_{12}}\|_1 \]
\[ \leq n \tau_2 R \lambda_n + (M_1 - s_0 \tau_1 \lambda_n)^{-1} n R s_0 \tau_2 \lambda_n^2 \]
\[ \leq 2 n \tau_2 R \lambda_n \]

and
\[ |w^T X_{(\gamma_2)} \Phi_{(\gamma_2)}| \leq \|w^T X_{(\gamma_2)} \| \|\beta_{(\gamma_2)}\|_1 \leq n \tau_2 R \lambda_n. \]

As a result,
\[ \frac{1}{n} |w^T (I - \Phi_{\gamma_1}) X_{(\gamma_2)} \Phi_{(\gamma_2)}| \leq \frac{1}{n} |w^T X_{(\gamma_2)} \Phi_{(\gamma_2)}| + \frac{1}{n} |w^T \Phi_{\gamma_1} X_{(\gamma_2)} \Phi_{(\gamma_2)}| \leq 3 \tau_2 R \lambda_n, \]
which concludes the second part with $L_2 = 3\tau_2 R/\sigma$.

**Condition 4.** Condition 4 can be derived directly using the inequality

$$\frac{1}{n} \lambda_{\min}(X_i^T X_i) \geq \min_i \frac{1}{n} \|x_i\|^2 - \sum_{j \neq i} \frac{1}{n} x_i^T x_j \geq M_1 - s_0 \tau_1 \lambda_n \geq M_1/2,$$

as long as $s_0 \tau_1 \lambda_n \leq M_1/2$.

Now combining Lemma 13 and 14, we immediately have Theorem 2.3 as stated below.

**Theorem C.2.** Assume all the conditions in Lemma 14 with $\lambda_n^2 = \sigma^2 \log p/n$ and the prior specification. If we have

$$\alpha > \frac{16\tau_2^2 + 26R^2\tau_2^2}{r'M_1} + 3 - \kappa, \quad A^2 \geq \frac{32}{M_1} (4q_1 + 7q_2 + 4r)(\alpha + \kappa + 5), \quad 12q_1 + 12r \leq q_2p^{2\alpha}$$

and

$$n \geq \max \left\{ \frac{8\tau_2^2 s_0}{r'M_1}, \frac{4s_0^2 \tau_1^2 \sigma^2}{M_1^2}, \frac{24\tau_2^2 R^2}{r'^2 \sigma^2} \right\} \log p,$$

then the posterior probability on the strong signals satisfies that

$$\pi_{n}(\gamma_\star | Y) \geq 1 - 3p^{-2}.$$
Bibliography


Biography

Xiangyu Wang was born on Jun 22, 1989 in Taizhou, Zhejiang, China. He received a B.S. in Mathematics from Peking University in July 2012, an M.S. in Computer Science from Duke University in 2016, and a Ph.D. in Statistics from Duke University in 2016.

He was the winner of the SBSS student travel award in JSM 2014. He also received student travel awards from the Neural Information Processing Systems (NIPS) conference in 2014 and 2015, and the International Conference on Machine Learning (ICML) in 2016.

His published papers include Wang and Leng (2015); Wang et al. (2015b, 2014); Wang and Dunson (2013); Wang et al. (2015c,a)