

SPARSE REGRESSION WITH EXACT CLUSTERING

A DISSERTATION  
SUBMITTED TO THE DEPARTMENT OF DEPARTMENT OF STATISTICS  
AND THE COMMITTEE ON GRADUATE STUDIES  
OF STANFORD UNIVERSITY  
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS  
FOR THE DEGREE OF  
DOCTOR OF PHILOSOPHY

Yiyuan She  
September 2008

© Copyright by Yiyuan She 2008  
All Rights Reserved

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

---

(Art Owen) Principal Adviser

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

---

(Trevor Hastie)

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

---

(Jerome Friedman)

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

---

(Guenther Walther)

Approved for the University Committee on Graduate Studies.

# Abstract

This dissertation deals with three closely related topics of the lasso in addition to supplying a comprehensive overview of the rapidly growing literature in this field.

The first part aims at improving the lasso to attain smaller prediction error while simultaneously keeping the model sparsity. We propose the data-augmented weighted lasso (DAWL) to make a natural combination of the lasso and other estimators like ridge regression. We investigate the data-augmentation starting from the ridge's nature in solving the singularity problem which successfully explains the reasonability of the elastic net, and from a non-asymptotic study of the lasso's variable selection which describes the roles of different parts of the Gram matrix played in lasso estimation and selection. A robust data-dependent scaling and a 'ranged lasso' are proposed to augment both the regression matrix (nondiagonally) and the response vector. In the discussions of weights, we prove a sharp oracle inequality for the weighted lasso in the orthogonal case, and propose  $z$ -value based weights with good asymptotics. Simulations show the advantages of DAWL in test error and sparsity.

The second topic is the study of a generic sparse regression problem with a customizable sparsity pattern matrix, motivated by, but not limited to, a supervised gene clustering problem in microarray data analysis. The 'clustered lasso' method is proposed with  $l_1$ -type penalties on both the coefficients and their pairwise differences. Somewhat surprisingly, it shows a quite different behavior than the lasso or the fused lasso; the granted power of the  $l_1$ -penalty to approximate the  $l_0$ -penalty seems specious in this situation. This leads us to a theoretical study of the power and limitations of the  $l_1$ -penalty in the general framework of sparse regression. We then discuss how to combine data-augmentation and weights to improve the naive  $l_1$ -penalty. To attack the challenging computation problem in high-dimensional space, we successfully generalize an iterative algorithm for solving the lasso and develop a novel accelerated 'annealing' algorithm with theoretical justifications. It applies

to any sparse regression like the fused/clustered lasso, and can handle a large design matrix as well as a large sparsity pattern matrix with apparent ease.

In the third part, we discuss a class of thresholding-based iterative selection procedures (TISP) for model selection and shrinkage. People have long before noticed the weakness of the convex  $l_1$ -constraint (or the soft-thresholding) in wavelets and have designed many different forms of nonconvex penalties to increase model sparsity and accuracy. But for a nonorthogonal regression matrix, there is great difficulty in both investigating the performance in theory and solving the problem in computation. TISP provides a simple and efficient way to tackle this so that we successfully borrow the rich results in the orthogonal design to solve the (nonconvexly) penalized regression for a general design matrix. Our starting point is, however, the thresholding rules rather than the penalty functions. Indeed, there is a universal connection between them. But a drawback of the latter is its non-unique form, and our way greatly facilitates the computation and the analysis. In fact, we are able to build the convergence theorem and explore the theoretical properties of the selection and the estimation via TISP nonasymptotically. More importantly, a novel Hybrid-TISP is proposed based on hard-thresholding and ridge-thresholding. It provides a fusion between the  $l_0$ -penalty and the  $l_2$ -penalty, and adaptively achieves the right balance between shrinkage and selection in statistical modeling. In practice, Hybrid-TISP shows superior performance in test-error and is parsimonious.

# Acknowledgements

First and foremost, I would like to thank my mentor Professor Art Owen for guiding me through the years. Art is a serious thinker and a practicalist. I must say his attitudes toward research have a far-reaching impact on me which has finally directed me toward an academic career. Art is also an excellent teacher and a man with great love and understanding. I am deeply indebted to him for dedicating so much time and effort to my technical reports, correcting my English writing with infinite patience, and supporting my family with kindness and generosity.

I would like to thank Professor Trevor Hastie for his extremely helpful suggestions and tremendous encouragement in my research and internship and job search during these years. I would also like to thank Professor Jerome Friedman for his insightful comments about my research and penetrating views. I am very grateful to Professor Guenther Walther for reading my thesis, and Professor Iain Johnstone and Professor Stephen Boyd for serving on my oral committee.

I would like to thank Bala Rajaratnam, who is not just a best friend but a brother to me. He gives me confidence and encouragement in both research and life. He has been a enormous help in my job search. I feel so proud of having such a talented and loving friend. Special thanks go to Hua Zhou for the many good times together in the day, and Ping Li for the fun conversations during the long dark nights.

Finally, I would also like to express my gratitude to my parents and my sister for love and support; in particular, I would like to thank my dear wife Jin Li, for her patience and understanding and cooking me delicious food every day. She is a gift from God and possesses all the virtues of a perfect woman. As I finish this thesis, we have our first baby, Jialai. As his names suggests, something beautiful and wonderful has come to my life again. I dedicate this thesis to both of them.

# Contents

<b>Abstract</b>	<b>iv</b>
<b>Acknowledgements</b>	<b>vi</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Lasso . . . . .	1
1.1.1 Theoretical explorations . . . . .	1
1.1.2 Computational algorithms . . . . .	3
1.1.3 Variants and improvements . . . . .	4
1.2 Outline . . . . .	5
<b>2 Improving Lasso: Data-augmentation and Weights</b>	<b>7</b>
2.1 Data Augmentation . . . . .	7
2.1.1 Two basic problems and ridge regression . . . . .	7
2.1.2 Lasso's variable selection . . . . .	9
2.1.3 Non-diagonal data augmentation . . . . .	13
2.1.4 Ranged lasso . . . . .	15
2.2 Weights . . . . .	16
2.3 Parameter Tuning and Numerical Examples . . . . .	22
2.3.1 Regularization parameter search . . . . .	23
2.3.2 Simulation results . . . . .	24
2.4 Discussion . . . . .	27
2.5 Proof of Theorem 1 . . . . .	28
2.6 Proof of Theorem 2 . . . . .	28
2.7 The Ranged Lasso Process . . . . .	35

2.8	Proof of Theorem 3 . . . . .	38
<b>3</b>	<b>Sparse Regression with Exact Clustering</b>	<b>40</b>
3.1	Background . . . . .	40
3.2	Clustered Lasso . . . . .	41
3.3	Limitations and Improvements of the Clustered Lasso . . . . .	45
3.3.1	The power and limitations of the $L_1$ -penalty in sparse regression . . . . .	45
3.3.2	Improving techniques . . . . .	49
3.3.3	Algorithm design and a simulation study . . . . .	52
3.4	A Fast Algorithm for Solving the Generic Sparsity Problem . . . . .	55
3.4.1	Motivation . . . . .	55
3.4.2	The ‘annealing’ algorithm . . . . .	58
3.4.3	Accelerated annealing . . . . .	61
3.4.4	Results on biological data . . . . .	66
3.5	Discussion . . . . .	67
3.6	Proofs of Proposition 1, Proposition 2, Theorem 4, and Theorem 5 . . . . .	68
3.7	Proofs of Theorem 6, Proposition 3, Theorem 7, Proposition 4, Proposition 5, and Proposition 6 . . . . .	74
<b>4</b>	<b>Thresholding-based Iterative Selection Procedures for Model Selection and Shrinkage</b>	<b>87</b>
4.1	Motivation . . . . .	87
4.1.1	From orthogonal designs to non-orthogonal designs . . . . .	87
4.2	TISP . . . . .	88
4.2.1	Thresholding rules and penalties . . . . .	88
4.2.2	TISP and its convergence . . . . .	90
4.3	Selection and Estimation via TISP . . . . .	93
4.4	TISP Designs and Numerical Examples . . . . .	96
4.4.1	An empirical study of TISPs . . . . .	96
4.4.2	Hybrid-TISP for model selection and shrinkage . . . . .	100
4.5	Discussion . . . . .	103
4.6	Proofs of Theorem 8, Proposition 8, and Proposition 9 . . . . .	104
4.7	Proof Outlines of Theorem 9, Theorem 10, Theorem 11, and Theorem 12 . . . . .	106
4.8	Proof of Theorem 13 . . . . .	106



# List of Tables

2.1	DAWL performance comparison . . . . .	26
3.1	Clustered lasso performance comparison . . . . .	55
4.1	TISP performance comparison . . . . .	99

# List of Figures

2.1	Ranged lasso process . . . . .	17
2.2	Lasso vs. weighted lasso . . . . .	18
2.3	SCAD, transformed L1, and weighted lasso. . . . .	19
3.1	Clustered lasso does not show enough exact-clustering effect . . . . .	45
3.2	Homogenous updating in AA . . . . .	62
3.3	Clustered coefficients from DAW-CLASSO . . . . .	67
4.1	Hybrid-penalty . . . . .	101

# Chapter 1

## Introduction

### 1.1 Lasso

Shrinkage methods are attractive in modeling and predictive learning because they offer continuous shrinkage with small generalization error, and they are usually easy to solve in practice. Among them, ridge regression and the lasso [50] are the most basic and popular ones. Both minimize a penalized likelihood function which is convex. Aside from the OLS fitting term, ridge regression imposes an  $l_2$ -norm constraint for all the variables, while the lasso (or basis pursuit [17] in the context of signal processing) uses an  $l_1$ -constraint.

The lasso draws people's particular attention because it provides an efficient and smooth way for variable selection, thereby achieving a sparse solution. Although in the orthonormal case it is well understood and has elegant theories [20, 5, 14], its shrinking and thresholding are not direct and clear enough for a general regression matrix, and it suffers some problems in selection and estimation [62, 15, 60]. In addition, the nonsmooth  $l_1$ -penalty poses challenges in both fast computation for high-dimensional data and theoretical analysis of its performance. There has been a large and rapidly growing body of literature for the lasso studies in the past three years. We classify these references into three categories and give a brief overview as follows.

#### 1.1.1 Theoretical explorations

Let  $\mathbf{y}$  be the response vector, and  $\mathbf{X}$  the design matrix of size  $n$ -by- $p$ . Suppose  $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ , and the noise is Gaussian, i.e.,  $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I})$ . We would like to recover  $\boldsymbol{\beta}$  under the sparsity assumption. It is no wonder that this variable selection problem, an  $l_0$ -norm optimization

indeed [19], is NP-hard [1]. The lasso amounts to solving

$$\min \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1,$$

where  $\lambda$  is the regularization parameter. Although it is intuitive that the  $l_1$ -norm provides a convex approximation to the  $l_0$ -norm and is itself nonsmooth at zero, it remains an interesting and meaningful topic in theory to what extent the lasso can help us identify the relevant variables and estimate their coefficients accurately.

Meinshausen & Bühlmann [37] and Zhao & Yu [60] are among the first to notice that the lasso selection is not always consistent, in contrast to traditional all-subset methods. The sufficient and the necessary conditions for sign-consistency are formalized as the irrepresentable conditions in [60]. Without going into the mathematical details, if the relevant predictors are highly correlated with the irrelevant ones, we can never expect the lasso to be sign consistent. Wainwright [53] obtains sharper bounds for the sparsity recovery probability under the (strong) irrepresentable condition.

The theoretical study of the lasso estimation was initiated by Knight & Fu [34], with  $p$  fixed and  $n \rightarrow \infty$ . Numerous asymptotic results then arise in this simple setting, e.g., Zou [61] and Yuan & Lin [57]; a useful notion here is the *oracle procedure* proposed by Fan & Li [26]. They are all referred to as fixed  $p$  discussions. Although this setup is classical and greatly simplifies the analysis, the “large  $p$ , small  $n$ ” case draws more and more attention because the lasso is a popular technique when  $p > n$ , say in the microarray data analysis.

Large- $p$  or nonasymptotic lasso theories are very technical. Candès *et al.* [15], Donoho *et al.* [19], in the context of signal processing, notice the importance of stable recovery for  $p > n$  and derive recovery bounds in a non-random setup (for noiseless signals and noisy ones). Bunea *et al.* [13] establish sparsity oracle inequalities for the  $l_1$  estimation-loss and the prediction-loss with large probability in a general nonparametric framework. These results are strengthened in Bickel *et al.* [8] where a simultaneous analysis of the lasso and Dantzig selector [16] is accomplished. Under certain conditions (weaker than the irrepresentable condition), Meinshausen & Yu [38] obtain the  $l_2$ -consistency of the lasso estimator. This indicates that although the lasso may not recover the correct sparsity pattern, under relaxed conditions, it still distinguishes between the truly zeros and the truly nonzeros with a separation zone, and all important variables (with sufficiently large coefficients) are selected with high probability. This point is supported by Zhang & Huang [58], where a deep

and systematic study regarding lasso's sparsity and bias is performed, and the *rate consistency* is eventually established for lasso. Nevertheless, all of these complicated theories are attained under some strong assumptions on the Gram matrix  $\mathbf{X}^T \mathbf{X}$ , such as the mutual coherence [19] and its variants [15], incoherent designs [38], the sparse Riesz condition [58], and the restrictive eigenvalue assumption [8]; in some sense, the design matrix cannot be too far from orthogonal to reach meaningful conclusions. These different assumptions make it difficult to compare the theoretical results with each other, and they also seem to be restrictive in applications.

### 1.1.2 Computational algorithms

Although the  $l_1$ -penalty results in a simple soft-thresholding for an orthogonal  $\mathbf{X}$ , the lasso solution has no closed form for a general design matrix. In the original lasso paper [50], Tibshirani used a quadratic programming method to solve it, after transforming the  $l_1$ -constraint into  $2^p$  linear constraints. Another popular idea is to treat the lasso as an iterative ridge regression and thus iterative reweighted least-squares (RWLS) can be used to get the estimate. Unfortunately, neither is fast enough in practice.

Later, more efficient procedures are proposed, including the homotopy method (Osborne *et al.* [42]), the well known LARS (Efron *et al.* [22]), and a recently re-discovered iterative algorithm (Friedman *et al.* [27], Wu & Lange [55]). Both of the first two exploit the piecewise linearity of the solution path. In particular, the least angle regression requires only  $p$  steps to get the entire set of solutions, the computation of which is of the same order as OLS. (Note that, however, the least angle solution path may not be identical to the lasso path; LAR solves a different yet similar problem.) We put more emphasis on the story of the last method which appears relatively new, but has, in fact, been proposed several times in different forms in history [28, 49, 18, 27, 55]. The procedure is amazingly simple, but surprisingly, according to Friedman, Hastie, & Tibshirani [27], it is “*... very competitive with the LARS algorithm, probably the fastest procedure for that (lasso) problem to-date*”.

Fu, in 1998, proposed the shooting procedure [28] for solving the lasso problem. His design is by taking the limit (theoretically) of the bridge estimator, which can be solved via a modified Newton-Raphson method, as its power index drops to 1. (However, the convergence proof (see Theorem 3 [28]) is incomplete.) The algorithm did not attract much attention at that time and was rediscovered (in vector form) in the field of mathematics in 2004 [18]; Daubechies *et al.* proved beautiful theoretical results on its convergence, covering

the large  $p$  case, in a general functional framework. It is worth mentioning that this theoretical achievement is considerably **stronger** than an ‘every accumulation point’ argument as we often seen in algorithm analysis (e.g., [7]). However, the authors did not pay much attention to its practical use (probably due to the fact that the convergence speed can be as slow as the Landweber iteration [24] for a small penalty value). Therefore, although this theoretical analysis makes a perfect practice of Browder’s and Opial’s classical works [12, 41] in pure mathematics, the algorithm was ignored, again, in the real world.

It is worth pointing out that, in statistics, we care about computing the *entire* solution path for tuning purposes, but not just a single solution for one  $\lambda$ -value. Hence a pathwise algorithm with warm start suits our needs and may facilitate the computation. Lately, Friedman *et al.* [27] and Wu & Lange [55] both derive this iteration from the point of view of coordinate optimization, and explicitly apply the pathwise algorithm to large data problems. They demonstrate its amazing performance in terms of the computation time compared to the homotopy method and LARS.

### 1.1.3 Variants and improvements

Various extensions and modifications to lasso have been proposed in the past three years. Some important variants in the statistics literature are the fused lasso [51], designed for problems with features that can be ordered, and the grouped lasso [56] and its CAP generalization [59], which can be used for naturally grouped predictors, such as the dummy variables introduced for a multi-level factor.

The Dantzig selector (DS) by Candès and Tao [16] is an important alternative to the lasso. The authors prove impressive oracle inequalities under the UUP condition. DS is computationally feasible since it reduces to a linear programming problem. However, its solution path is often jittery and does not show better performance than the lasso [23].

There have been many different ways suggested to improve the lasso in both variable selection and prediction accuracy. For example, to combine the strengths of lasso and ridge, Zou & Hastie propose the elastic net [62], which adopts a linear combination of the  $l_1$ -penalty and the  $l_2$ -penalty, while Owen designs a novel convex ‘Berhu’ penalty [43] with concomitant scale estimation to replace the naïve  $l_1$ .

As an example of the two-stage methods, the adaptive lasso by Zou [61] uses the OLS estimate to construct a weighted  $l_1$ -penalty in fitting the model to bring more sparsity. The

nonnegative garotte [11] is closely related to it. A recent advance is the one-step SCAD [63] with the weights constructed from the OLS estimate via nonconvex penalty functions. Although in the fixed  $p$  setting the adaptive lasso is an oracle procedure asymptotically, its weight construction in the large- $p$  situation remains a challenging problem both in theoretical analysis and for practical use.

Noticing that the lasso utilizes the same parameter for model selection and shrinkage, Meinshausen proposes the relaxed lasso [36] which introduces two tuning parameters to control the amount of shrinkage and the selection. It takes the plain lasso and the LARS-OLS hybrid [22] as two extremes. A further improvement is the VISA proposed by Radchenko & James [44].

## 1.2 Outline

Chapter 2 discusses data augmentation and weights to improve the lasso by making a natural fusion of the lasso and any other estimate. One motivation for data-augmentation is from ridge regression which solves the singularity problem, and this starting point successfully explains the reasonability of the elastic net and significantly generalizes it. Another motivation is from our nonasymptotic results about how the lasso can solve the sparsity problem, which describes the roles of different parts of the Gram matrix played in estimation risks and sign consistencies. To augment the regression matrix and the response vector, we introduce a safe data-dependent scaling against poor estimates, and a ‘ranged lasso’ with consistently small test errors. In the discussions of weights, we prove sharp oracle inequalities for the weighted lasso in the orthogonal case, develop scale invariant estimation procedures, and propose a  $z$ -value based weight construction with good asymptotics. Simulations show the advantages of this data-augmented weighted lasso (DAWL) in test error and sparsity.

Chapter 3 studies a generic sparse regression problem with a customizable sparsity pattern matrix, motivated by, but not limited to, a supervised gene clustering problem in microarray data analysis. The ‘clustered lasso’ method is proposed with  $l_1$ -type penalties on both the coefficients and their pairwise differences. Somewhat surprisingly, it shows a quite different behavior than the lasso or the fused lasso; the granted power of the  $l_1$ -penalty to approximate the  $l_0$ -penalty seems specious in this situation. This leads us to an asymptotic study of the power and limitations of the  $l_1$ -penalty in sparse regression. We then discuss

how to combine data-augmentation and weights to improve the naïve  $l_1$ -penalty.

In the second part, we consider a computation problem. Recently, an iterative algorithm, simple but fast, was rediscovered to solve the lasso. We successfully generalize it both in practice and in theory and propose an ‘annealing’ algorithm which applies to the generic sparse regression (taking the lasso and the fused lasso as special cases). Some effective accelerating techniques are further investigated to boost the convergence. The accelerated annealing (AA) algorithm, involving only matrix multiplication and thresholding, can handle a large design matrix as well as a large sparsity pattern matrix with apparent ease.

Chapter 4 discusses a class of thresholding-based iterative selection procedures (TISP) for model selection and shrinkage. People have long before noticed the weakness of the convex  $l_1$ -constraint (or the soft-thresholding) in wavelets and have designed many different forms of nonconvex penalties to increase model sparsity and accuracy. But for a nonorthogonal regression matrix, there is great difficulty in both investigating the performance in theory and solving the problem in computation. TISP provides a simple and efficient way to tackle this so that we successfully borrow the rich results in the orthogonal design to solve the (nonconvexly) penalized regression for a general design matrix.

Our starting point is, however, the thresholding rules rather than the penalty functions. Indeed, there is a universal connection between them. But a drawback of the latter is its non-unique form, and our way greatly facilitates the computation and the analysis. In fact, we are able to build the convergence theorem and explore the theoretical properties of the selection and the estimation via TISP nonasymptotically.

More importantly, a novel Hybrid-TISP is proposed based on hard-thresholding and ridge-thresholding. It provides a fusion between the  $l_0$ -penalty and the  $l_2$ -penalty, and adaptively achieves the right balance between shrinkage and selection in statistical modeling. In practice, Hybrid-TISP shows superior performance in test-error and is parsimonious.

Chapter 2, Chapter 3, and Chapter 4 are based on three technical reports [45, 46, 47], respectively.

## Chapter 2

# Improving Lasso: Data-augmentation and Weights

This chapter aims at improving the lasso to attain smaller predictor error while simultaneously keeping the model sparsity. We propose the data-augmented weighted lasso (DAWL) to make a natural combination of the lasso and other estimators like ridge regression. Section 2.1 investigates the data-augmentation, starting from the ridge's nature in solving the singularity problem, and from a non-asymptotic study of the lasso's variable selection. We introduce a data-dependent scaling and a 'ranged lasso' to do the augmentation in both the regression matrix and the response vector. Section 2.2 discusses the weights. We prove a sharp oracle inequality of the weighted lasso in the orthonormal case, and study different weight choices, where a  $z$ -value based weight construction is proposed with good asymptotics. In Section 2.3, a practical and important issue is addressed – the multiple regularization parameter search; simulation results are also reported. The technical details are left to the end.

### 2.1 Data Augmentation

#### 2.1.1 Two basic problems and ridge regression

Shrinkage methods, such as ridge regression and the lasso [50], are widely used in statistical estimation. Generally, they help solve two different types of problems in regression: (a) the

*Singularity* problem and (b) the *Sparsity* problem.

The singularity problem occurs when the  $n$ -by- $d$  input matrix  $\mathbf{X}$  does not have full column rank (numerically), especially when  $n < d$ , whereas the sparsity problem is posed by the additional requirement of the solution to be sparse, which is often meaningful both in practice and in theory when the dimensionality is high.

The lasso is mainly intended to solve the sparsity problem. Although it can sometimes alleviate the singularity (when  $n < d$ ), the lasso can not guarantee to remove all the singularity, say, caused by highly correlated columns of  $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_d]$ . This is exactly the starting point of the elastic net [62] (or eNet, for short, in this chapter).

Ridge regression, on the other hand, perfectly solves the singularity problem. In fact, if seen as a way to overcome the singularity problem, the ridge can be thought of as a *data augmentation* technique, and it does not have to shrink all coefficients towards 0. We use  $\mathbf{X} \rightarrow \mathbf{y}$  to denote the regression problem with  $\mathbf{X}$  as the input matrix, and  $\mathbf{y}$  as the response vector. Then the ridge regression (or the modified ridge regression), instead of solving  $\mathbf{X} \rightarrow \mathbf{y}$ , considers the problem of

$$\begin{bmatrix} \mathbf{X} \\ \sqrt{\lambda}\mathbf{I} \end{bmatrix} \rightarrow \begin{bmatrix} \mathbf{y} \\ \times \end{bmatrix}. \quad (2.1)$$

$\sqrt{\lambda}\mathbf{I}$  helps to decorrelate the columns of the design matrix, and thus solves the singularity problem. The ‘ $\times$ ’ part is not necessarily  $\mathbf{0}$ , although it is often so. For example, a simple choice is  $\sqrt{\lambda}\hat{\boldsymbol{\beta}}_{indv}$  with  $\hat{\boldsymbol{\beta}}_{indv} = [\mathbf{x}_i^T \mathbf{y} / (\mathbf{x}_i^T \mathbf{x}_i)]_{d \times 1}$ , the univariate estimate. In fact, following this idea, we can give the eNet (the corrected version) [62] a simple interpretation, one that explains why an extra factor should come in compared to the naïve eNet.

In [62], noticing the weakness of the lasso in solving the singularity problem, Zou and Hastie therefore introduced the *naïve* eNet as a combination of the lasso and ridge regression by imposing both an  $l_1$  penalty and an  $l_2$  penalty. Furthermore, to guard against *double shrinkage*, they designed an empirical way to improve the naïve eNet estimate by multiplying it by a factor of  $1 + \lambda_2$  (see Section 3.2 of [62] for more details). The final eNet estimate, according to Theorem 2 of [62], is defined to be

$$\hat{\boldsymbol{\beta}}_{\lambda_1, \lambda_2}^{(eNet)} = \arg \min_{\boldsymbol{\beta}} \boldsymbol{\beta}^T \frac{\mathbf{X}^T \mathbf{X} + \lambda_2 \mathbf{I}}{1 + \lambda_2} \boldsymbol{\beta} - 2\mathbf{y}^T \mathbf{X} \boldsymbol{\beta} + \lambda_1 \|\boldsymbol{\beta}\|_1 \quad (2.2)$$

for given  $\lambda_1, \lambda_2$ , assuming  $\mathbf{x}_i^T \mathbf{x}_i = 1$  for  $i = 1, \dots, d$ . We can prove the following result.

**Theorem 1** *Given  $\lambda_1, \lambda_2 > 0$ , define*

$$\hat{\boldsymbol{\beta}}_{\lambda_1, \lambda_2} = \arg \min \left\| \begin{bmatrix} \mathbf{y} \\ \sqrt{\lambda_2} \mathbf{X}^T \mathbf{y} \end{bmatrix} - \begin{bmatrix} \mathbf{X} \\ \sqrt{\lambda_2} \mathbf{I} \end{bmatrix} \boldsymbol{\beta} \right\|_2^2 + \lambda_1 \|\boldsymbol{\beta}\|_1. \quad (2.3)$$

Then  $\hat{\boldsymbol{\beta}}_{\lambda_1, \lambda_2} = \hat{\boldsymbol{\beta}}_{\frac{\lambda_1}{1+\lambda_2}, \lambda_2}^{(eNet)}$ .

See Section 2.5 for the proof.

Moreover, since the eNet used a search strategy of first choosing a grid of values for  $\lambda_2$ , then searching over the  $\lambda_1$ -space for any fixed  $\lambda_2$ , our best  $\hat{\boldsymbol{\beta}}_{\lambda_1^*, \lambda_2^*}$  based on (2.3) gives exactly the eNet estimate. In short, the eNet solves the lasso regression of

$$\begin{bmatrix} \mathbf{X} \\ \sqrt{\lambda} \mathbf{I} \end{bmatrix} \rightarrow \begin{bmatrix} \mathbf{y} \\ \sqrt{\lambda} \hat{\boldsymbol{\beta}}_{indv} \end{bmatrix} \quad (2.4)$$

with  $\hat{\boldsymbol{\beta}}_{indv} = [\mathbf{x}_i^T \mathbf{y} / (\mathbf{x}_i^T \mathbf{x}_i)]_{d \times 1}$ , the univariate estimate. Accordingly, the design of the eNet can be viewed as a special case of applying the data augmentation technique to solve the singularity problem.

### 2.1.2 Lasso's variable selection

In addition to the above study of ridge regression and the singularity problem, there is another important motivation for data augmentation based on how the lasso can solve the sparsity problem, which is often neglected. The lasso is known to be a convex *approximation* of  $l_0$ -constrained variable selection problem. But it is legitimate to ask when this is true and how good the approximation is. To get a taste of its performance, we give a simple nonasymptotic theorem.

Given the input matrix  $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_d]$ , observe  $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$  with  $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2]$ ,  $\boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{bmatrix}$ ,  $\boldsymbol{\beta}_1 = \mathbf{0}$ , and  $\boldsymbol{\epsilon} \sim N(0, \sigma^2 \mathbf{I})$ , where  $\mathbf{X}, \mathbf{X}_1, \mathbf{X}_2$  are of size  $n \times d, n \times d_1, n \times d_2$ , and  $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2$  are of size  $d_1 \times 1, d_2 \times 1$ , respectively, and  $n, d_1, d_2 \geq 1$ . The goal is to estimate

the sparse  $\boldsymbol{\beta}$ . Let

$$\boldsymbol{\Sigma} = \mathbf{X}^T \mathbf{X} = \begin{bmatrix} \mathbf{X}_1^T \mathbf{X}_1 & \mathbf{X}_1^T \mathbf{X}_2 \\ \mathbf{X}_2^T \mathbf{X}_1 & \mathbf{X}_2^T \mathbf{X}_2 \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Sigma}_1 & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{12}^T & \boldsymbol{\Sigma}_2 \end{bmatrix}.$$

Assume for the moment that  $\mathbf{X}$  has been column-normalized, that is,  $\mathbf{x}_i^T \mathbf{x}_i = 1$  for  $i = 1, \dots, d$ , or  $\boldsymbol{\Sigma}$ 's diagonal entries are all 1s. Introduce the sign function for any  $\mathbf{v} \in \mathfrak{R}^d$ :  $\text{sgn}(\mathbf{v}) = [\text{sgn}(v_i)]_{d \times 1}$  with  $\text{sgn}(v_i) = 1, -1$ , or  $0$  when  $v_i > 0, < 0$ , or  $= 0$ , respectively.

Under these settings, we consider the performance of the lasso estimate  $\hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\boldsymbol{\beta}}_1 \\ \hat{\boldsymbol{\beta}}_2 \end{bmatrix} = \arg \min \frac{1}{2} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1$  for any  $\lambda > 0$ :

**Theorem 2** Define  $\tau_1, \tau_2$  as the smallest eigenvalues of  $\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2$ , respectively, and  $\kappa \triangleq \max_{1 \leq i \leq d_1} \|\mathbf{v}_i\|_2 / \sqrt{d_2}$  for  $\boldsymbol{\Sigma}_{12} = [\mathbf{v}_1, \dots, \mathbf{v}_{d_1}]^T$ , to control the magnitude of the entries in  $\boldsymbol{\Sigma}_{12}$ . Assume  $\tau_2 > 0, \tau_2 \geq d_2 \kappa$ .

1. *Sign Consistency.* First,

$$P(\hat{\boldsymbol{\beta}}_1 \neq \mathbf{0}) \leq 1 - (1 - 2\Phi(-M))^{d_1} \leq 2d_1 \cdot \frac{1}{M} \varphi(M), \quad (2.5)$$

where  $M = \frac{1}{\sigma} \left(1 - \frac{\kappa d_2}{\tau_2}\right) \lambda$ ,  $\varphi(\cdot)$  is the standard normal density, i.e.,  $\varphi(M) = \frac{1}{\sqrt{2\pi}} e^{-\frac{M^2}{2}}$ . Moreover, assume  $\beta_{2,i} \neq 0, \forall i = 1, \dots, d_2$ , and  $L \triangleq \frac{1}{\sigma} \left(\sqrt{\tau_2} \cdot \min |\boldsymbol{\beta}_2| - \lambda \frac{d_2}{\sqrt{\tau_2}}\right)$  is nonnegative, then

$$P(\text{sgn}(\hat{\boldsymbol{\beta}}) = \text{sgn}(\boldsymbol{\beta})) \geq [1 - 2\Phi(-M)]^{d_1} [1 - 2\Phi(-L)]^{d_2}, \quad (2.6)$$

where  $\Phi$  is the standard normal distribution.

2. *Risks.* Denote the individual risks of  $\hat{\boldsymbol{\beta}}_1$  and  $\hat{\boldsymbol{\beta}}_2$  by  $R_1$  and  $R_2$ , respectively, i.e.,  $R_1 = E\|\hat{\boldsymbol{\beta}}_1 - \mathbf{0}\|_2^2$ , and  $R_2 = E\|\hat{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_2\|_2^2$ . Assume  $\boldsymbol{\Sigma}$  is nonsingular (and so  $n \geq d_1 + d_2$ ). Then

$$R_2 \leq 3 \left[ \frac{d_2}{\tau_2} \sigma^2 + \lambda^2 \frac{d_2}{\tau_2^2} + \kappa^2 \frac{d_1 d_2}{\tau_2^2} \cdot R_1 \right]. \quad (2.7)$$

On the other hand,

$$R_1 \leq \frac{\sigma^2 d_1^2}{\tau_1^2} (K_1 M + K_2 \frac{1}{M}) \varphi(M), \quad (2.8)$$

$$\text{with } K_1 = 6 \cdot \frac{1 + \frac{1 + \kappa^2 d_2^2 / \tau_2^2}{(1 - \kappa d_2 / \tau_2)^2}}{(1 - \kappa^2 \frac{d_1 d_2}{\tau_1 \tau_2})^2}, \quad K_2 = 6 \left(1 - \kappa^2 \frac{d_1 d_2}{\tau_1 \tau_2}\right)^{-2}, \text{ in which we assume } \kappa^2 \cdot \frac{d_1 d_2}{\tau_1 \tau_2} \leq 1.$$

See Section 2.6 for its proof.

This theorem describes the performance of the lasso by giving explicit bounds for any finite  $(n, d_1, d_2)$ , and states the *different* roles of  $\Sigma_1$ ,  $\Sigma_2$ , and  $\Sigma_{12}$  in terms of  $\tau_1$ ,  $\tau_2$ , and  $\kappa$ , respectively. We can see that both  $R_1$  and  $R_2$  are monotonically increasing functions of  $1/\tau_1, 1/\tau_2$ , and  $\kappa$ . And the size of  $\kappa$  is especially important.

There are many theoretical results [34, 15, 19, 53, 60, 38, 61] regarding the lasso's performance. Our motivation for developing this theorem in a *nonasymptotic* sense is to identify important and intuitive factors in the lasso selection and find ways to improve it. To assure the lasso is effective, we should have, at the minimum, good control over  $P(\hat{\beta}_1 \neq \mathbf{0})$  and  $R_2$ . Reducing  $\kappa$  to have  $\mathbf{X}_1 \perp \mathbf{X}_2$  approximately true is one way. Indeed, a very small  $\kappa$  can help reduce  $R_1$  and  $R_2$ , and weaken the interference of  $\mathbf{X}_1(\Sigma_1)$  in  $R_2$  or  $\mathbf{X}_2(\Sigma_2)$  in  $R_1$ , such that the lasso risks  $R_1$  and  $R_2$  are primarily determined by  $\tau_1$  and  $\tau_2$  separately, in addition to an extra  $\lambda$ -contribution. As an interesting example, consider  $d_i = 4$ ,

$$\Sigma = \begin{bmatrix} 0.99(\mathbf{1} \cdot \mathbf{1}^T) + 0.01\mathbf{I} & 0.01(\mathbf{1} \cdot \mathbf{1}^T) \\ 0.01(\mathbf{1} \cdot \mathbf{1}^T) & 0.1(\mathbf{1} \cdot \mathbf{1}^T) + 0.9\mathbf{I} \end{bmatrix}_{8 \times 8}, \quad (2.9)$$

and  $\beta = [0, 0, 0, 0, 3, 3, 3, 3]^T$ . The lasso still works well here since  $\kappa$  is pretty small.

Our theorem assumes  $\mathbf{x}_i^T \mathbf{x}_i = 1$  for  $i = 1, \dots, d$  for notational simplicity, yet it conceals the role of the sample size  $n$ . In fact, if  $\text{diag}(\mathbf{X}^T \mathbf{X}) \leq \sigma_{max}^2$ , that is, the  $l_2$ -norms of columns of  $\mathbf{X}$  are no greater than  $\sigma_{max}$ , we only need to replace the  $\beta$ ,  $\hat{\beta}$ ,  $\lambda$ , and  $\mathbf{X}$  in Theorem 2 by  $\beta \cdot \sigma_{max}$ ,  $\hat{\beta} \cdot \sigma_{max}$ ,  $\lambda/\sigma_{max}$ , and  $\mathbf{X}/\sigma_{max}$  respectively. For example, if  $\mathbf{X}$  is normalized to have column  $l_2$ -norms equal to  $\sqrt{n}$ , which is often convenient in the asymptotic setting of  $n \rightarrow \infty$ , then for  $\hat{\beta} = \arg \min \frac{1}{2} \|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \lambda \|\beta\|_1$ ,  $\Sigma$  should be defined as  $\frac{1}{n} \mathbf{X}^T \mathbf{X}$ ,

$\tau_1, \tau_2, \kappa$  are defined in the same way as before, but

$$M = \frac{1}{\sigma} \left( 1 - \frac{\kappa d_2}{\tau_2} \right) \frac{\lambda}{\sqrt{n}}, \quad L = \frac{1}{\sigma} \left( \sqrt{\tau_2} \cdot \min |\boldsymbol{\beta}_2| \cdot \sqrt{n} - \frac{\lambda}{\sqrt{n}} \frac{d_2}{\sqrt{\tau_2}} \right), \quad (2.10)$$

and (2.7), (2.8) become

$$R_2 \leq \frac{3}{n} \left[ \frac{d_2}{\tau_2} \sigma^2 + \frac{\lambda^2 d_2}{n \tau_2^2} + \kappa^2 \frac{d_1 d_2}{\tau_2^2} \cdot n R_1 \right], \quad \text{and} \quad R_1 \leq \frac{\sigma^2 d_1^2}{n \tau_1^2} (K_1 M + K_2 \frac{1}{M}) \varphi(M). \quad (2.11)$$

In this way, we clearly see the contribution of the sample size  $n$ .

A suitable value of  $\lambda$  should be chosen: it should not be too small from the bound for  $R_1$ , and not too large, either, seen from  $R_2$ ; the same is true concerning the sign consistency bound (2.6). In fact, as a corollary, we have

$$P(\text{sgn}(\hat{\boldsymbol{\beta}}) \neq \text{sgn}(\boldsymbol{\beta})) \leq 2d_1 \frac{1}{M} \varphi(M) + 2d_2 \frac{1}{L} \varphi(L). \quad (2.12)$$

We can develop asymptotic results for  $\lambda$ . Assume  $\mathbf{X}$  is normalized such that its column  $l_2$ -norms equal  $\sqrt{n}$ . Consider two special cases. (i) Suppose  $\frac{1}{n} \mathbf{X}^T \mathbf{X} \rightarrow \mathbf{C}$  as  $n \rightarrow \infty$ , but  $d_i, \boldsymbol{\beta}$  are fixed. Then under some regularity conditions we get: if  $\lambda(n)/\sqrt{n} \rightarrow \infty$  and  $\lambda(n)/n \rightarrow 0$ , then the lasso is sign consistent. This result coincides with other studies like [34, 60]. Moreover, we know that under this condition,  $R_1 \rightarrow 0, R_2 \rightarrow 0$  by (2.11). (ii) Let  $d_2, \boldsymbol{\beta}_2$  fixed,  $n(\iota), d_1(\iota) \rightarrow \infty$  as  $\iota \rightarrow \infty$ , and  $1 - \kappa(\iota) d_2 / \tau_2(\iota) \geq$  some positive constant for  $\iota$  large enough. Then  $\hat{\boldsymbol{\beta}}$  is sign consistent if  $d_1 \varphi(M)/M \rightarrow 0$  and  $\lambda/n \rightarrow 0$ , which only requires  $n$  to grow faster than  $\log d_1$ ! It also follows from the risk bound (2.11) that if  $\boldsymbol{\Sigma}(\iota)$  is nonsingular and  $\tau_1(\iota) \geq c > 0$  for  $\iota$  large enough, (and so  $n \geq d$ ), then  $M = \sqrt{2 \log \frac{d_1^2}{n} + (1 + \epsilon) \log \log \frac{d_1^2}{n}} \sim \sqrt{2 \log \frac{d_1^2}{n}} \leq \sqrt{2 \log d_1}$  (for any  $\epsilon > 0$ ) is sufficient to ensure  $R_1 \rightarrow 0$  under the regularity conditions stated in the theorem. From Donoho and Johnstone's work [20] in the orthogonal design, this also implies this risk bound can not be improved significantly in general.

Regarding the sign consistencies, we find that our result is most similar to Wainwright's Proposition 1 [53], but more intuitive, and sharper in some sense: for example, in the second situation (ii), our first condition says that  $M = \sqrt{c \log d_1}$  for any  $c \geq 2$  is enough, while [53] requires  $M^2 / \log d_1 \rightarrow \infty$ . The latter is incapable of giving the right rate, say, in the asymptotic study of the optimal choice of  $\lambda$  in wavelets [20]. Our non-asymptotic sign

consistency results bound the  $p$ -values. For example, we all know that the lasso sets some  $\hat{\beta}_i$ 's to be exactly zero, say,  $\hat{\beta}_1 = \mathbf{0}$ , but we are also curious to learn and should report how strong the belief is as in hypothesis testing; (2.5) provides an easy-to-compute bound for the  $p$ -value of  $H_0 : \beta_1 = 0$  if we get  $\hat{\beta}_1 = \mathbf{0}$ .

The risk results are new, although the classical work in the wavelet or orthonormal case [20] about the risk oracle inequalities and the minimax optimality (in an asymptotic sense) of  $\lambda = \sqrt{2 \log n}$  is well known. Candès *et al.* [15], Donoho *et al.* [19] derived recovery bounds in a *non-random* setup — for either noiseless signals or noisy ones — where  $\|\epsilon\|_2$  is bounded, mainly by use of a mutual coherence  $M \triangleq \max_{i \neq j} |\Sigma_{ij}|$ , or some other variants. Mutual coherence can be related to the minimum eigenvalue of  $\Sigma$ :  $\lambda_{\min}(\Sigma) \geq 1 - (d-1)M$ ; this needs  $M \leq 1/(d-1)$  to be meaningful, which is sometimes too restrictive for  $M$ . And the analysis by a *global* mutual coherence is more or less rough and can not differentiate between the roles of the different parts in  $\Sigma$ . Consider the previous example (2.9) again: the lasso estimate is good with a suitable  $\lambda$  though the mutual coherence is very high (0.99) and  $\Sigma_1$  is nearly singular.

In summary, to ensure that the lasso plays a good role in variable selection, we must reduce  $\kappa$ . Again, **data augmentation**, say in the diagonal manner (2.1) with an additional  $\sqrt{\lambda}\mathbf{I}$  below  $\mathbf{X}$ , is an effective way. (Of course, assuming no prior information of the positions of zeros in  $\beta$ , we have to decorrelate the whole matrix  $\mathbf{X}$ .) And by Theorem 1, the widely accepted eNet is a special case. The problem is then to figure out what to put in the augmented  $\mathbf{X}$ , and more importantly, what to put in the augmented  $\mathbf{y}$  since the true  $\beta$  is unknown.

### 2.1.3 Augment $\mathbf{X}$ : Non-diagonal augmentation using data-dependent scaling

First let's study data augmentation in the eNet:  $\begin{bmatrix} \mathbf{X} \\ \sqrt{\lambda}\mathbf{I} \end{bmatrix} \rightarrow \begin{bmatrix} \mathbf{y} \\ \sqrt{\lambda}\hat{\beta}_{indv} \end{bmatrix}$ , where  $\hat{\beta}_{indv} = [\mathbf{x}_i^T \mathbf{y} / (\mathbf{x}_i^T \mathbf{x}_i)]_{d \times 1}$ . Ignoring the  $l_1$ -constraint for the moment, we consider

$$\hat{\beta} = \arg \min \left\| \begin{bmatrix} \mathbf{y} \\ \sqrt{\lambda}\hat{\beta}_{indv} \end{bmatrix} - \begin{bmatrix} \mathbf{X} \\ \sqrt{\lambda}\mathbf{I} \end{bmatrix} \beta \right\|_2^2.$$

By the SVD of  $\mathbf{X}$ :  $\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^T$ , the prediction  $\hat{\mathbf{y}} \triangleq \mathbf{X}\hat{\boldsymbol{\beta}}$  can be shown to be

$$\hat{\mathbf{y}} = (1 + \lambda) \cdot \mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y} = \sum_{i=1}^d \frac{d_i^2(1 + \lambda)}{d_i^2 + \lambda} (\mathbf{u}_i^T\mathbf{y})\mathbf{u}_i, \quad (2.13)$$

where  $\mathbf{u} = [\mathbf{u}_1, \dots, \mathbf{u}_d]$ ,  $\mathbf{D} = \text{diag}\{d_i\}_{i=1}^d$ . (In comparison,  $\hat{\mathbf{y}}_{ols} = \sum(\mathbf{u}_i^T\mathbf{y})\mathbf{u}_i$ .) So when  $d_i < 1$ , the projection of  $\mathbf{y}$  on  $\mathbf{u}_i$  is shrunk; when  $d_i > 1$ , the projection on  $\mathbf{u}_i$  is extended. The threshold value ‘1’ here is not data-dependent; we’d rather replace it by a scale parameter. One way is to solve the following problem jointly

$$\min_{(\boldsymbol{\beta}, s)} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta} - s\hat{\boldsymbol{\beta}}_{indv}\|_2^2 \quad \text{s.t.} \quad \|\boldsymbol{\beta}\|_1 \leq c,$$

or equivalently (optimizing over  $s$ ),

$$\min \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \left\| \boldsymbol{\beta} - \frac{\sum \hat{\boldsymbol{\beta}}_{indv,i}\boldsymbol{\beta}_i}{\sum \hat{\boldsymbol{\beta}}_{indv,i}^2} \hat{\boldsymbol{\beta}}_{indv} \right\|_2^2 \quad \text{s.t.} \quad \|\boldsymbol{\beta}\|_1 \leq c.$$

Introduce  $\mathbf{Z}(\boldsymbol{\beta})$  to be a matrix constructed from the nonzero vector  $\boldsymbol{\beta}$  in the following way

$$\mathbf{Z}(\boldsymbol{\beta}) \triangleq \mathbf{I} - \boldsymbol{\beta}\boldsymbol{\beta}^T / \|\boldsymbol{\beta}\|_2^2.$$

Then the above optimization can be rewritten as

$$\min \left\| \begin{bmatrix} \mathbf{y} \\ \mathbf{0} \end{bmatrix} - \begin{bmatrix} \mathbf{X} \\ \sqrt{\lambda} \cdot \mathbf{Z}(\hat{\boldsymbol{\beta}}_{indv}) \end{bmatrix} \boldsymbol{\beta} \right\|_2^2 \quad \text{s.t.} \quad \|\boldsymbol{\beta}\|_1 \leq c. \quad (2.14)$$

In fact, assuming  $\mathbf{X}$  has been column normalized, we know

$$\hat{\boldsymbol{\beta}}_{indv} = \mathbf{X}^T\mathbf{y} = (\mathbf{X}^T\mathbf{X})\boldsymbol{\beta} + \mathbf{X}^T\boldsymbol{\epsilon} = \boldsymbol{\Sigma}\boldsymbol{\beta} + \mathbf{X}^T\boldsymbol{\epsilon},$$

which is not even consistent in the nonorthogonal case, so the introduction of  $s$  is necessary and helpful, say, in the following example,

$$\boldsymbol{\Sigma} = 0.5 (\mathbf{1} \cdot \mathbf{1}^T + \text{diag}\{\mathbf{1}\})_{40 \times 40}, \quad \text{and} \quad \boldsymbol{\beta}^T = \begin{bmatrix} 2 & \dots & 2 & 0 & \dots & 0 \end{bmatrix}$$

as Test 4 in [50]. To the best of our knowledge, this *non-diagonal* way of data-augmentation

(2.14) is novel. And the test results in Section 2.3 show an improvement brought by this scale parameter.

It is worth pointing out that in decorrelating the columns of the regression matrix,  $\mathbf{Z}(\hat{\boldsymbol{\beta}}_{indv}) = \mathbf{I} - \hat{\boldsymbol{\beta}}_{indv}\hat{\boldsymbol{\beta}}_{indv}^T/\|\hat{\boldsymbol{\beta}}_{indv}\|_2^2$  is *singular*: its eigenvalues are 1 with multiplicity  $d - 1$ , and 0 with multiplicity 1; the eigenvector corresponding to 0 is  $\hat{\boldsymbol{\beta}}_{indv}/\|\hat{\boldsymbol{\beta}}_{indv}\|_2$ . However, the whole input matrix

$$\begin{bmatrix} \mathbf{X} \\ \sqrt{\lambda}\mathbf{Z} \end{bmatrix}$$

is rank-deficient if and only if the vector  $\mathbf{X}\hat{\boldsymbol{\beta}}_{indv}$  is exactly  $\mathbf{0}$ . So generally speaking, the data augmentation of (2.14) still helps to solve the singularity problem. (The same is true for a general  $\hat{\boldsymbol{\beta}}_{aug}$  in (2.15) below.) And it is easy to see this data-dependent way is more robust since we do not blindly use all the information provided by  $\boldsymbol{\beta}_{indv}$  do the data-augmentation — in fact, we do not have to because of the tuning from the regularization parameter  $\lambda$ , and this one degree of freedom is saved in the construction of  $\mathbf{Z}$ .

#### 2.1.4 Augment $\mathbf{y}$ : Ranged lasso

In general, we can replace  $\hat{\boldsymbol{\beta}}_{indv}$  by some more accurate estimate  $\hat{\boldsymbol{\beta}}_{aug}$  (such as  $\hat{\boldsymbol{\beta}}_{ridge}$ ) in the following way for data augmentation:

$$\min \left\| \begin{bmatrix} \mathbf{y} \\ \mathbf{0} \end{bmatrix} - \begin{bmatrix} \mathbf{X} \\ \sqrt{\lambda}\mathbf{Z}(\hat{\boldsymbol{\beta}}_{aug}) \end{bmatrix} \boldsymbol{\beta} \right\|_2^2 \quad \text{s.t.} \quad \|\boldsymbol{\beta}\|_1 \leq c. \quad (2.15)$$

Hereinafter, denote the augmented  $\mathbf{y}$  and the augmented  $\mathbf{X}$  by  $\tilde{\mathbf{y}}$  (or  $\tilde{\mathbf{y}}_\lambda$ ),  $\tilde{\mathbf{X}}$  (or  $\tilde{\mathbf{X}}_\lambda$ ), respectively. With no knowledge of how accurate  $\hat{\boldsymbol{\beta}}_{aug}$  is, we still do the scaling because it is a safe and robust choice against poor estimates.

A typical choice of  $\hat{\boldsymbol{\beta}}_{aug}$  might be the ridge estimate  $\hat{\boldsymbol{\beta}}_{ridge}$ . However, according to our experience,  $\hat{\boldsymbol{\beta}}_{ridge}$  is not really a good estimate all the time. For example, in the experiments carried out in Section 2.3, ridge regression hardly achieves a small test error. Another disadvantage is  $\hat{\boldsymbol{\beta}}_{ridge}$  is not sparse at all, and it may seriously hurt the sparsity of the final estimate. Correspondingly, an estimate with consistently small test error (and better some sparsity) is preferred in augmenting the data. In our experiment, a ‘ranged

lasso' estimate is used:

$$\hat{\boldsymbol{\beta}}_{r-lasso} = \arg \left( \min \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1 \text{ s.t. } \max \boldsymbol{\beta} - \min \boldsymbol{\beta} \leq c \right) \quad (2.16)$$

where  $\max \boldsymbol{\beta} \triangleq \max_{1 \leq i \leq d} \beta_i$ ,  $\min \boldsymbol{\beta} \triangleq \min_{1 \leq i \leq d} \beta_i$ . The range constraint, sort of like ridge regression, limits the expansion of the estimate.

In the orthogonal case  $\mathbf{X}^T \mathbf{X} = \mathbf{I}$ , where we may write  $\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2$  as  $\|\boldsymbol{\beta} - \boldsymbol{\beta}_{obs}\|_2^2$  in (2.16) with  $\boldsymbol{\beta}_{obs}$  known, by the KKT conditions, evaluating  $\hat{\boldsymbol{\beta}}_{r-lasso}$  is a *combined* process of the lasso and  $\max \boldsymbol{\beta} / \min \boldsymbol{\beta}$  auto-clustering (see Figure 2.1): (i) It does the lasso shrinkage and thresholding first on  $\boldsymbol{\beta}_{obs}$ , the result denoted by  $\hat{\boldsymbol{\beta}}_{lasso}$ ; (ii) Given bounds  $a \leq b$ , set all  $\hat{\beta}_{lasso,i}$  less than  $a$  to be  $a$ , and all  $\hat{\beta}_{lasso,i}$  greater than  $b$  to be  $b$  — denote the total amount of change by  $\Delta_1(a)$  and  $\Delta_2(b)$ , respectively; (iii) Starting from  $a = -\infty, b = +\infty$ , move  $a, b$  towards each other until  $\Delta_1(a) = \Delta_2(b)$  and the range constraint is satisfied. Section 2.7 gives a KKT argument for this. In the non-orthogonal case, the clustering and the thresholding are still present, but can not be clearly separated, and the range constraint has a complex shrinking effect. An empirical search strategy of how to choose the two regularization parameters in the ranged lasso is given in Section 2.3.

Surprisingly, the ranged lasso often shows superior performance in test error even in the case of true  $\boldsymbol{\beta}$  not having big  $\max/\min$  clusters, compared to the usual estimates (lasso, ridge, DS [16], and also eNet). Although the solution is sometimes not sparse enough, it at least offers a competitive alternative to the ridge estimate in some sense.

Replacing  $\hat{\boldsymbol{\beta}}_{indv}$  by a more accurate estimate like  $\hat{\boldsymbol{\beta}}_{r-lasso}$  is a promising way to attain smaller test error than eNet, but it is also more expensive in general. Finally we point out that our data augmentation technique (like (2.14) or (2.15)) applies to the situation  $n < d$ .

## 2.2 Weights

The data-augmented lasso offers a natural way to fuse  $\hat{\boldsymbol{\beta}}_{lasso}$  and  $\hat{\boldsymbol{\beta}}_{aug}$  (such as  $\hat{\boldsymbol{\beta}}_{ridge}$ ). But it can hurt the sparsity sometimes (due to the fact that we often use a not-so-sparse  $\hat{\boldsymbol{\beta}}_{aug}$ ). That the sign consistency results of Theorem 2 have nothing to do with  $\tau_1$  suggests another way, aside from data augmentation, to reduce  $\kappa$ : rescale the columns of  $\mathbf{X}$  before carrying out the lasso. Imagine we could replace  $\mathbf{X}$  by  $\mathbf{X}\mathbf{D} = \begin{bmatrix} \mathbf{X}_1 & \mathbf{X}_2 \end{bmatrix} \cdot \begin{bmatrix} \mathbf{D}_1 & \\ & \mathbf{D}_2 \end{bmatrix}$  with  $\mathbf{D}_1 = \text{diag}\{\mathbf{0}\} = \mathbf{0}$ ,  $\mathbf{D}_2 = \text{diag}\{\mathbf{1}\} = \mathbf{I}$ . Then  $\boldsymbol{\Sigma}_{12}$  would become  $\mathbf{D}_1 \mathbf{X}_1^T \mathbf{X}_2 \mathbf{D}_2$ , and

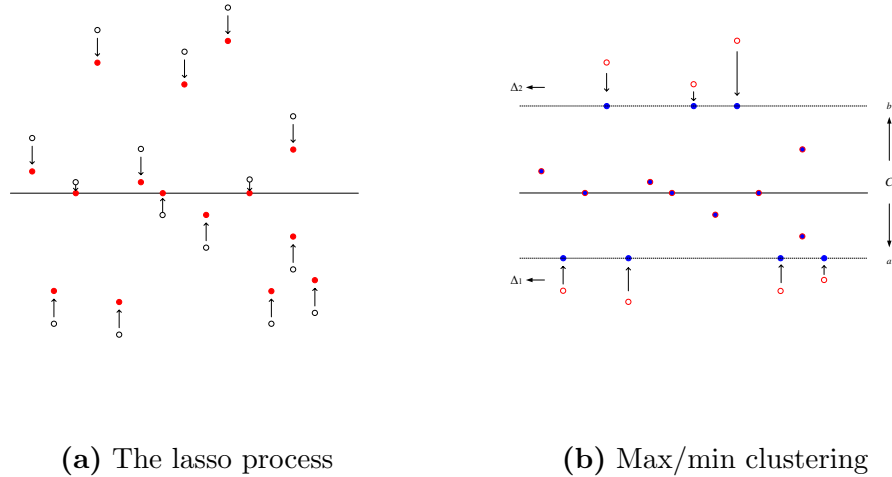


Figure 2.1: Ranged lasso process in the orthogonal case. The first stage, as shown in (a), is the lasso process, in which the observed  $\beta_{obs}$  (black circles) is moved to  $\hat{\beta}_{lasso}$  (red dots). Then, in (b), max/min clustering is carried out on  $\hat{\beta}_{lasso}$  (red circles) to get the final ranged lasso estimate  $\hat{\beta}_{r-lasso}$  (blue dots); in this stage, the upper bound  $b$  and the lower bound  $a$  determine the amount of changes at two ends, denoted by  $\Delta_1$  and  $\Delta_2$ , respectively. And we have  $\Delta_1 = \Delta_2$  and  $b - a = c$  in the end.

$\kappa$  would be 0, indicating that we would have good control over  $P(\hat{\beta}_1 \neq \mathbf{0})$  and  $R_2$  (even if  $\tau_1$  would be 0).

Let's consider a naïve example with  $\mathbf{X} = \mathbf{I}_{2 \times 2}$ ,  $\beta = [\beta_1, \beta_2]^T$ ,  $\beta_1 > 0$ ,  $\beta_2 = 0$ , and  $\mathbf{y} = \mathbf{X}\beta + \epsilon$ . See Figure 2.2. In Figure 2.2a, we use a simple  $l_1$  constraint, corresponding to some  $l_1$ -balls that are square. But considering the position of the observation  $\mathbf{y}$ , we may want to try parallelograms instead as shown by those transformed polytopes in Figure 2.2b.

It is clear that the transformed polytopes in Figure 2.2b, by borrowing some information from  $\mathbf{y}$ , make more sense than the  $l_1$ -balls in (2a). It helps us get an estimate with both sparsity and good fit. In this section, we shall consider the *weighted* lasso as follows, with weights  $\{w_i\}$  constructed from *both*  $\mathbf{X}$  and  $\mathbf{y}$ :

$$\min \|\mathbf{y} - \mathbf{X}\beta\|_2^2 \quad \text{s.t.} \quad \sum w_i |\beta_i| \leq c. \quad (2.17)$$

We notice that this is exactly the same idea Zou proposed in [61], where he calls it the

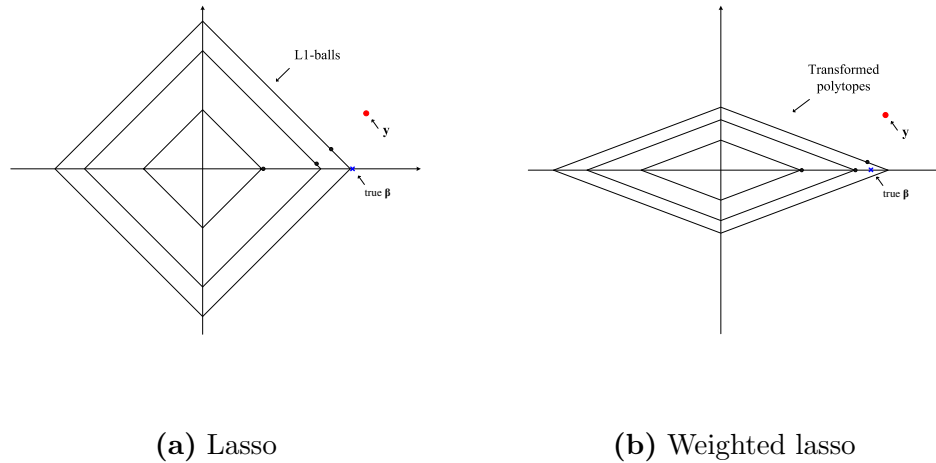


Figure 2.2: The lasso vs. the weighted lasso. This is a naïve example in two-dimensional space, with  $\mathbf{y}$  (big red dots) being an observation of  $\beta$  (blue crosses), contaminated with noise. We give three examples of the constraint value, corresponding to three L1-balls and three transformed polytopes in (a) and (b), respectively. So an estimate (small black dot) would be the point closest to  $\mathbf{y}$  that lies in the corresponding squared L1-ball or parallelogram. It is clear that the estimates in (b) are sparser and more accurate than those in (a).

adaptive lasso and uses the OLS estimate construct the weights.

The weights  $\{w_i\}$  are not something unfamiliar to us. In fact, the usual way of normalizing the columns of  $\mathbf{X}$  before applying the lasso is equivalent to (2.17) with

$$w_i = \sqrt{\mathbf{x}_i^T \mathbf{x}_i} \triangleq \sigma_i. \tag{2.18}$$

(This  $w_i$  does not depend on  $\mathbf{y}$ .) The weighted lasso can be attractive because a set of *good* weights has the advantages of (i) increasing the sparsity of  $\hat{\beta}$ , and (ii) reducing the biases for large  $\hat{\beta}_i$ 's.

Take the orthogonal case for an illustration. Suppose  $\mathbf{X}^T \mathbf{X} = \mathbf{I}$ , and so  $\hat{\beta}_{indv} = \mathbf{X}^T \mathbf{y}$ . Set  $w_i = 1/|\hat{\beta}_{indv,i}|^\eta$  with  $\eta \geq 0$ . Then the solution to  $\min \frac{1}{2} \|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \lambda \sum w_i |\beta_i|$  is

$$\hat{\beta}_i = \hat{\beta}_{indv,i} - \frac{\lambda}{|\hat{\beta}_{indv,i}|^\eta} \text{sgn}(\hat{\beta}_i), \tag{2.19}$$

where  $\widetilde{\text{sgn}}$  is a simple notation — for the solution  $\hat{\beta}$ ,  $\widetilde{\text{sgn}}(\hat{\beta}_i)$  takes the value 1,  $-1$ , or  $\hat{\beta}_{\text{indv},i}/(\lambda|\hat{\beta}_{\text{indv},i}|^{-\eta})$  when  $\hat{\beta}_i > 0$ ,  $< 0$ , or  $= 0$ , respectively (see Section 2.6 for details). (2.19) reduces the bias of the lasso estimate, and takes soft- and hard-thresholding as two extremes (corresponding to  $\eta = 0, \infty$ , respectively).

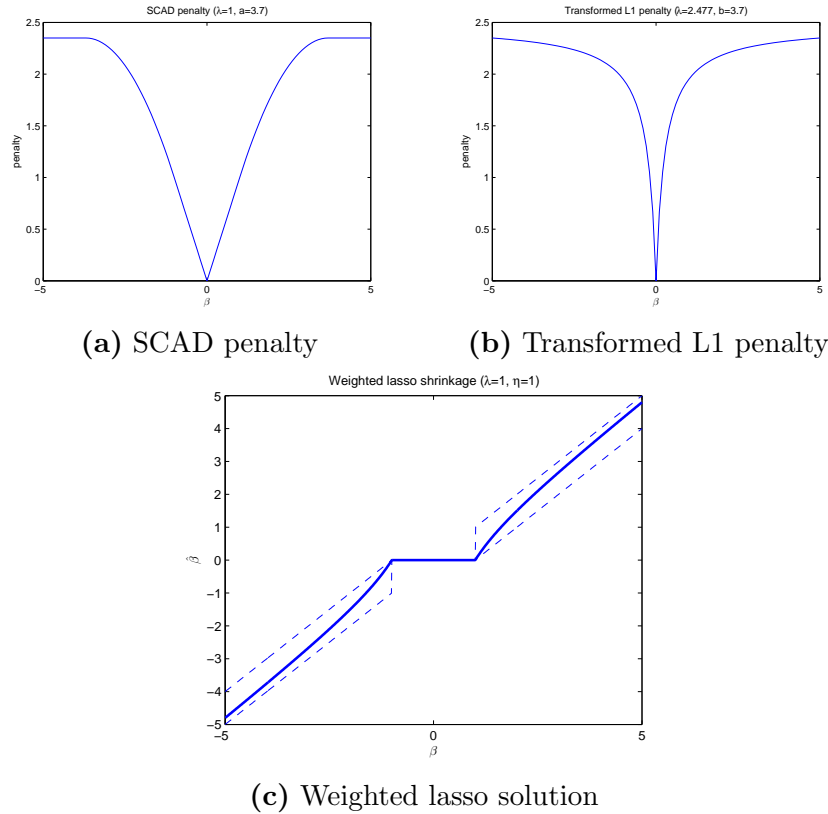


Figure 2.3: SCAD penalty, the transformed L1 penalty, and the weighted lasso solution in one-dimensional space.

The well known SCAD [5] and the transformed  $L_1$ -penalty [39] also take soft- and hard-thresholding as two extremes, by use of concave penalties (see Figure 2.3a, 2.3b). The weighted lasso achieves the same goal (Figure 2.3c), with no need to resort to local optimization like the ROSE [5]. It is convex, and as simple as the lasso. Note that thresholdings determined by (2.19) are not uniform in all coordinates, but the critical points  $\pm\lambda^{\frac{1}{\eta+1}}$  are the same.

Motivated by the above simple facts, we can get the oracle inequality for the weighted lasso estimate  $\hat{\beta}$  in (2.19).

**Theorem 3** Let  $\lambda_0 = \lambda^{1+\eta}/\sigma$ . Then

$$E\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2^2 \leq (1 + \lambda_0^2) \cdot \sum_1^n \min\left(\frac{2\varphi(\lambda_0)}{\lambda_0}\sigma^2 + \beta_i^2, \sigma^2\right) \quad (2.20)$$

for any  $\lambda_0 > 1$ . Consequently, when  $\lambda_0 = \sqrt{2\log n}$ ,

$$E\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2^2 \leq (2\log n + 1) \left( \frac{\sigma^2}{\sqrt{\pi\log n}} + \sum \min(\beta_i^2, \sigma^2) \right) \quad (2.21)$$

for any  $n \geq 2$ .

As a matter of fact, (2.20) holds for *any* estimator  $\hat{\boldsymbol{\beta}}$  bounded by the soft- and hard-thresholding estimates with threshold value  $\lambda_0\sigma$ . A proof is provided in Section 2.8, where we strengthen Donoho and Johnstone's risk result for hard-thresholding in the nonasymptotic situation. A comparison with the oracle bound obtained by Zou [61] is given at the end of this section.

Next we consider the case when  $\mathbf{X}$  is not orthogonal. First define the *scale invariant* property for a given estimating procedure  $P$ , or an estimate  $\hat{\boldsymbol{\beta}} = P(\mathbf{X}, \mathbf{y})$ : If  $\mathbf{X}$  is replaced by  $\mathbf{X} \cdot \mathbf{S}$ , where  $\mathbf{S}$  is any diagonal matrix  $\text{diag}\{s_i\}_{i=1}^d$  with  $s_i \neq 0$  (but can be negative), then the estimate  $\hat{\boldsymbol{\beta}}$  changes to  $\mathbf{S}^{-1}\hat{\boldsymbol{\beta}}$ . It is easy to verify that  $\hat{\boldsymbol{\beta}}_{\text{indv}}$ ,  $\hat{\boldsymbol{\beta}}_{\text{ols}}$  (if it takes a consistent form when it is not unique), and the lasso (2.17) with  $w_i \propto \sigma_i$  are such examples. But how can one construct a scale invariant procedure in general? Suppose we have an estimate  $\hat{\boldsymbol{\beta}}_{\text{wts}}$  for our problem  $\mathbf{X} \rightarrow \mathbf{y}$  that is scale invariant. Then we see that for any  $\eta \geq 0$ , the weighted lasso (2.17) with

$$w_i \propto |\hat{\beta}_{\text{wts},i}|^{-\eta} \sigma_i^{1-\eta} \quad (2.22)$$

is also scale invariant. Moreover, under the assumption, it is easy to show this weighted lasso can be obtained from the following implementation: (i) Column-normalize  $\mathbf{X}$  to get a new design matrix  $\mathbf{X}' = \mathbf{X}\mathbf{S}$ , with its column  $l_2$ -norms equal to each other; here, the diagonal entries of  $\mathbf{S}$  are required to be positive; (ii) construct the weights  $\{w'_i\}$  for the regression problem  $\mathbf{X}' \rightarrow \mathbf{y}$ , the corresponding weighted lasso estimate denoted by  $\hat{\boldsymbol{\beta}}'$ ; (iii) let  $\hat{\boldsymbol{\beta}} = \mathbf{S}\hat{\boldsymbol{\beta}}'$ . Clearly,  $w'_i \propto |\hat{\beta}'_{\text{wts},i}|^{-\eta}$  corresponds to the weights constructed in (2.22). As a matter of fact, it is not hard to see that this 3-step procedure is always scale invariant as

long as  $w_i$  are constructed from the regression problem of  $\mathbf{X}' \rightarrow \mathbf{y}$ . We adopt this scaling procedure for the lasso problem, and thus only need to focus on the second step, which can be combined with step (i) in implementation.

The choice of  $\hat{\beta}_{wts}$  is crucial to applying the weighted lasso in the non-orthogonal situation. Our experiences show that the weight technique has limited improvement in test error unless  $\hat{\beta}_{wts}$  is exceptionally accurate.

Now combining weights with the data augmentation technique, we obtain the general form of data-augmented weighted lasso (DAWL). For the weighted version of (2.15), i.e.,

$$\min \left\| \begin{bmatrix} \mathbf{y} \\ \mathbf{0} \end{bmatrix} - \begin{bmatrix} \mathbf{X} \\ \sqrt{\lambda} \mathbf{Z}(\hat{\beta}_{aug}) \end{bmatrix} \beta \right\|_2^2 \quad \text{s.t.} \quad \sum w_i |\beta_i| \leq c, \quad (2.23)$$

we may put  $\hat{\beta}_{wts}$  to be  $\hat{\beta}_{aug}$ , or the OLS estimate of the augmented problem  $\tilde{\mathbf{X}}_\lambda \rightarrow \tilde{\mathbf{y}}_\lambda$  (remember that  $\tilde{\mathbf{X}}_\lambda$  has full rank in general). In all of our experiments, the power index  $\eta$  in (2.22) is fixed at 1.

Sometimes the penalization according to the size of  $|\hat{\beta}_{wts,i}|$  is not enough: if  $var(\hat{\beta}_{wts,i})$  is relatively large, we have good reason to shrink  $\hat{\beta}_i$  to benefit from the bias-variance trade-off. To incorporate this standard error information, we may use the  $z$ -values  $\frac{\hat{\beta}_{wts,i}}{\widehat{se}(\hat{\beta}_{wts,i})}$ , or  $\frac{\hat{\beta}_{wts,i}}{\sqrt{n} \widehat{se}(\hat{\beta}_{wts,i})} \triangleq z_i$  to construct  $w_i$  as an alternative of  $\hat{\beta}_{wts,i}$ , if the standard errors  $\widehat{se}(\hat{\beta}_{wts,i})$  are available (for nonzero  $\hat{\beta}_{wts,i}$ ). It makes sense as  $z_i$  is a statistical measurement of how confident  $\hat{\beta}_{wts,i}$  is from 0, taking its randomness into consideration. This way penalizes based on the significance of  $\hat{\beta}_{wts}$ , but not its absolute value. It shrinks  $\hat{\beta}_i$  if we meet a relatively small  $\hat{\beta}_{wts,i}$ , or a large  $var(\hat{\beta}_{wts,i})$ , or both.

Since the  $w_i$ 's are now dependent on both  $\mathbf{X}$  and  $\mathbf{y}$ , they are random, and nonasymptotic bounds like Theorem 2 are more difficult to obtain. Recently, Zou [61] shows (i) good asymptotic properties of it with  $d_i$  fixed and  $n \rightarrow \infty$  (the classic setup), and (ii) oracle inequalities in the orthonormal case. Due to an error made in the derivative calculation of the weighted estimator, his oracle bound in Theorem 3 should be

$$(2 \log n + 5 + 4\eta) \cdot \left( \sum \min(\beta_i^2, \sigma^2) + \frac{\sigma^2}{2\sqrt{\pi \log n}} \right),$$

with the first factor being  $(2 \log n + 5 + 4\eta)$  instead of  $(2 \log n + 5 + 4/\eta)$ , which diverges as  $\eta$  goes to infinity. This bound is weaker than ours as shown in Theorem 3. This is because Stein’s lemma does not handle well the oracle bound for an estimator very close to hard thresholding, since the hard-thresholding function is *not* weakly differentiable. (See the comments at the end of Section 2.8 for details.)

In the same asymptotic setup with only  $n$  growing to infinity (see Section 2 of [61] for details), we can show that, although  $\mathbf{z}$  is not a  $\sqrt{n}$ -consistent estimator to  $\boldsymbol{\beta}$ , the  $z$ -value based weights  $w_i = |z_i|^{-\eta}$  enjoy exactly the same *oracle properties* of Theorem 2 in [61], because Zou’s Theorem 2 can be easily generalized to any sequence  $\mathbf{z}(n)$  satisfying:

$$a_n(\mathbf{z}(n) - \boldsymbol{\gamma}) = O_p(1),$$

where  $a_n \rightarrow \infty$ ,  $\boldsymbol{\gamma}$  is a constant vector with  $|\text{sgn}(\boldsymbol{\gamma})| = |\text{sgn}(\boldsymbol{\beta})|$ ; and  $w_i = 1/f(|z_i|)$  where  $f(t) = t^\eta (\eta \geq 0)$ ,  $te^t$ , and  $a^{\frac{1}{t}} (0 < a < 1)$ , are all valid choices. (The proof follows the same lines as [61] and is omitted. A much more general result regarding weights is given in Chapter 3, or see She [46].)

This result implies that  $\mathbf{z}(n)$  does *not* need to be an estimator of  $\boldsymbol{\beta}$ . It can be used to justify the use of  $z$ -values in weight construction, from an asymptotic point of view, and provides a far more general way for weight construction.

## 2.3 Parameter Tuning and Numerical Examples

We did simulations to test the performance of the data-augmented weighted lasso (DAWL). But before introducing our results, we must address a very important problem – the multiple regularization parameter search, which is inescapable to do regularization well in implementation. Theoretically we should do a multi-dimensional grid search for all possible combinations of regularization parameters, but quite often we can not afford this in the real world. (Even if we could, there would still be difficulty in choosing the ‘best’ one since there is a trade-off between test error and sparsity.) In this section, we look for an efficient and effective empirical search strategy to get us a good solution with both small test error and some sparsity.<sup>1</sup>

---

<sup>1</sup>We treat the test error as a more important criterion than the sparsity in this chapter.

### 2.3.1 Regularization parameter search

Either the problem of data-augmented lasso (2.15), or its weighted version (2.23), or the ranged lasso (2.16) is involved with *two* regularization parameters. So an effective search strategy is necessary in implementation. In fact, in the case of multiple regularization parameters, the **parameter search** seems to count a lot in determining an algorithm's practical performance at least as much as, if not more than, the **penalty type**. Without a proper search scheme, we cannot make a sincere judgement on an algorithm's power. We discuss the problem in this subsection and give an empirical *alternate* search after reviewing existing approaches. In the following discussions, we assume there is always a validation dataset available besides the training data.

The fused lasso [51] and the eNet also have two regularization parameters, and the same search problem. In the eNet, the parameter search is done in the following way: (i) pick a small grid of values for one parameter; (ii) fix this parameter at any of these values, and search over the entire path indexed by the other parameter; (iii) choose the best combination of the two parameters which gives the smallest validation error. Similarly, the fused lasso picked a grid (un-evenly spaced) in the first step based on the approximately computed degrees of freedom. In (ii), instead of fixing one and searching for the other, it moves along a line in the parameter space in the direction  $(1, -2)$ , which is made empirically (see [51] for more details). Both the fused lasso and the eNet make use of the efficient LARS [22] algorithm in generating the lasso path.

The difficulty of applying the eNet search is how to choose a small grid for the first parameter which covers the potentially good values. In [62], the grid was made to be  $[0, 0.01, 0.1, 1, 10, 100]$ . It works for eNet most of the time because  $\hat{\beta}_{indv}$  is often *not* a good estimate and the grid happens to focus on smaller values, which is, nevertheless, not necessarily true for a general  $\hat{\beta}_{aug}$  in DAWL.

In tuning the regularization parameters for our problems, we do an *alternative* search for two *properly* chosen regularization parameters. For example, for the ranged lasso problem (2.16), instead of searching in the  $(\lambda, c)$  space, we write it in the form of

$$\min \|\beta\|_1 + \eta(\max \beta - \min \beta) \quad \text{s.t.} \quad \|\mathbf{y} - \mathbf{X}\beta\|_2^2 \leq \delta,$$

and choose  $(\eta, \delta)$  to be the regularization parameters to do the search. Note in this way, the

error term and the sparsity penalty are separated, with one parameter in each. In the first step, setting  $\eta = 1$ , we generate and search over the first path with respect to  $\delta$  to find the best solution (with the smallest validation error) at, say,  $\delta^{(o)}$ . Next, fixing  $\delta$  at  $\delta^{(o)}$ , search along the  $\eta$ -path to get an optimal value  $\eta^{(o)}$  with the smallest validation error. Then,  $\eta$  is fixed again, this time at  $\eta^{(o)}$ , to do the search over the  $\delta$ -path. Finally, we compare the results from the 3 searches, and take  $(\eta, \delta)$  to be the one minimizing the validation error.

Similarly, for the weighted version of the data-augmented lasso (2.23), which can be put as  $\min \|\tilde{\mathbf{y}}_\lambda - \tilde{\mathbf{X}}_\lambda \boldsymbol{\beta}\|_2^2$  s.t.  $\sum w_i |\beta_i| \leq c$  (see (2.15) for the definitions of  $\tilde{\mathbf{X}}_\lambda, \tilde{\mathbf{y}}_\lambda$ ), we search in the  $(\lambda, c)$  space, and adopt the same search scheme, but run it twice with two initial values 1, 50 (empirically chosen) for  $\lambda$ . So six paths are generated and searched in total. In the end we compare the results from them to get one with the minimum validation error.

### 2.3.2 Simulation results

We did experiments on 5 simulation datasets. Each contains training data, validation data, and test data. The last three are from examples 1, 2, 4 in [50], or the first 3 examples in [62], except that the validation datasize is increased to 100 in examples 3, 4 below.<sup>2</sup> As [62], we use  $\# = \text{“} \cdot / \cdot / \cdot \text{”}$  to denote the number of observations in the training data, the validation data, and the test data. Let  $\boldsymbol{\Sigma}$  be the correlation matrix in generating  $\mathbf{X}$ , i.e., each row of  $\mathbf{X}$  is independently drawn from  $N(\mathbf{0}, \boldsymbol{\Sigma})$ . We use  $(\{a_1\}^{n_1}, \dots, \{a_k\}^{n_k})$  to denote the column vector made by  $n_1$   $a_1$ 's,  $\dots$ ,  $n_k$   $a_k$ 's consecutively in the following examples.

**Example 1.**  $\# = 50/100/200$ ,  $d = 32$ ,  $\boldsymbol{\beta} = (\{0\}^{10}, \{4\}^5, \{1\}^5, \{-2\}^5, \{-0.5\}^7)$ ,  $\sigma = 5$ ,  $\Sigma_{ij} = \rho^{|i-j|}$  with  $\rho = 0.5$ .

**Example 2.**  $\# = 40/80/80$ ,  $d = 14$ ,  $\boldsymbol{\beta} = (\{4\}^1, \{2\}^1, \{1\}^2, \{0\}^5, \{-1.5\}^1, \{-3\}^1, \{-4\}^1, \{-6\}^1, \{-12\}^1)$ ,  $\sigma = 8$ ,  $\Sigma_{ij} = \rho^{\min(\alpha_{ij}^{(1)}, \alpha_{ij}^{(2)}, \alpha_{ij}^{(3)}, \alpha_{ij}^{(4)})}$ , where  $\alpha_{ij}^{(1)} = \max(i, j)$ ,  $\alpha_{ij}^{(2)} = \max(d + 1 - i, j)$ ,  $\alpha_{ij}^{(3)} = \max(i, d + 1 - j)$ ,  $\alpha_{ij}^{(4)} = \max(d + 1 - i, d + 1 - j)$ , and  $\rho = 0.75$  for  $i \neq j$ .

**Example 3.**  $\# = 20/100/200$ ,  $d = 8$ ,  $\boldsymbol{\beta} = (\{3\}^1, \{1.5\}^1, \{0\}^2, \{2\}^1, \{0\}^3)$ ,  $\sigma = 3$ ,  $\Sigma_{ij} = \rho^{|i-j|}$  with  $\rho = 0.5$ .

**Example 4.**  $\# = 20/100/200$ ,  $d = 8$ ,  $\boldsymbol{\beta} = (\{3\}^1, \{1.5\}^1, \{0\}^2, \{2\}^1, \{0\}^3)$ ,  $\sigma = 3$ ,  $\Sigma_{ij} = \rho^{|i-j|}$  with  $\rho = 0.85$ .

**Example 5.**  $\# = 100/100/400$ ,  $d = 40$ ,  $\boldsymbol{\beta} = (\{0\}^{10}, \{2\}^{10}, \{0\}^{10}, \{2\}^{10})$ ,  $\sigma = 15$ ,  $\Sigma_{ij} = 0.5$

<sup>2</sup>To guard against misleading judgements of a method's performance caused by unwisely chosen values of the regularization parameters.

for  $i \neq j$ .

Before an algorithm is applied, the columns of a regression matrix are all normalized to have squared  $l_2$ -norm equal to the number of the observations; no centering is done in these examples.

Each model is simulated 50 times, and then we measured the performance of each method by fitness and sparsity. The fitness is characterized by the 40% trimmed-mean<sup>3</sup> of the scaled MSE (SMSE) on the test data, where SMSE is  $100 \cdot (\sum_{i=1}^N (\hat{y}_i - y_i)^2 / (N\sigma^2) - 1)$  defined for the test data. And the sparsity here is the mean number of zeros in the estimates.

Thirteen methods in total were implemented in Matlab and their results are listed in Table 2.1. Let's look at the first 5 ones in the table. Owen's Berhu penalty [43] brings sparsity and its test error can be as small as ridge's. Dantzig selector (DS) [16] is comparable to the lasso, but not significantly better. The ranged lasso consistently shows excellent performance in test error for all datasets (including Example 2 which does not have max/min clusters at all) although it may not be very sparse sometimes, e.g. Example 5. It offers a competitive alternative among the first five estimates.

We also implemented 7 versions of the data-augmented weighted lasso (DAWL) to improve the lasso by fusing the lasso and some other estimate. In the first two,  $\hat{\beta}_{indv}$  is used as  $\hat{\beta}_{aug}$ , while the ranged lasso estimate  $\hat{\beta}_{r-lasso}$  is employed in the rest 5. The first, eNet\*, is our eNet implementation using the alternative search. But as mentioned,  $\hat{\beta}_{indv}$  may be a poor estimate in the nonorthonormal case, and DAWL.2 shows the strength of the non-diagonal augmentation using a data-dependent scaling introduced in Subsection 2.1.3. DAWL.3 can give even smaller test error by using the more accurate ranged lasso estimate to do the data augmentation.

DAWL.4–7 test some weights. DAWL.4 and DAWL.5 construct the weights via the ranged lasso and ridge regression, respectively. DAWL.6 and DAWL.7 make use of the augmented OLS estimate; observe that the latter way of  $z$ -value based weights, which corresponds to significance penalization, is indeed a feasible choice. Drawing a comparison between DAWL.3 and these weighted approaches, we learn that the weight technique, though showing limited power in test error, improves the sparsity.

In particular, let's study an interesting dataset – Example 5. If  $\hat{\beta}_{aug}$  is neither good nor sparse, like  $\hat{\beta}_{indv}$  in eNet or DAWL.1 for this dataset, the improvement in test error is still

---

<sup>3</sup>Medians of errors are used mostly [50, 62] to measure the performance from multiple runs, but are not so stable for comparisons from our experience. Discarding 20 highest and 20 lowest errors, we compute the average of the remaining 10.

Table 2.1: Performance comparisons between different algorithms on simulation data 1–5, in terms of test error (40% trimmed-mean SMSE) and sparsity (mean # of zeros in the estimates).

	Example 1		Example 2		Example 3		Example 4		Example 5	
	Test-err	Sparsity	Test-Err	Sparsity	Test-err	Sparsity	Test-err	Sparsity	Test-err	Sparsity
Ridge	<b>65.6</b>	0.0	<b>37.4</b>	0.0	<b>27.9</b>	0.0	<b>20.4</b>	0.0	<b>9.9</b>	0.0
Berhu	<b>63.4</b>	6.5	<b>35.6</b>	1.9	<b>28.2</b>	1.7	<b>18.8</b>	2.0	<b>11.1</b>	8.4
Lasso	<b>68.4</b>	11.0	<b>30.0</b>	3.8	<b>27.5</b>	3.0	<b>19.7</b>	3.7	<b>19.9</b>	18.9
DS	<b>62.9</b>	9.8	<b>30.8</b>	4.0	<b>26.2</b>	2.9	<b>18.1</b>	3.5	<b>23.0</b>	19.4
Ranged Lasso	<b>50.5</b>	9.2	<b>29.0</b>	3.5	<b>21.6</b>	2.0	<b>15.8</b>	2.2	<b>10.2</b>	5.5
eNet	<b>54.7</b>	9.4	<b>30.4</b>	4.2	<b>23.4</b>	2.6	<b>14.3</b>	2.6	<b>15.1</b>	15.0
DAWL.1 (eNet*)	<b>53.8</b>	9.8	<b>29.1</b>	4.3	<b>23.2</b>	2.5	<b>14.2</b>	2.3	<b>14.8</b>	15.3
DAWL.2	<b>50.3</b>	6.1	<b>29.3</b>	3.7	<b>23.7</b>	2.0	<b>15.4</b>	2.1	<b>10.7</b>	7.2
DAWL.3	<b>49.2</b>	5.8	<b>26.7</b>	2.9	<b>19.8</b>	1.9	<b>14.6</b>	1.9	<b>9.5</b>	4.9
DAWL.4	<b>51.6</b>	7.9	<b>27.3</b>	3.2	<b>19.6</b>	3.4	<b>15.5</b>	3.4	<b>10.3</b>	8.7
DAWL.5	<b>48.8</b>	7.5	<b>24.8</b>	4.0	<b>20.4</b>	3.1	<b>15.8</b>	2.6	<b>9.9</b>	8.1
DAWL.6	<b>49.5</b>	8.5	<b>25.4</b>	4.5	<b>18.8</b>	3.4	<b>16.0</b>	3.2	<b>9.8</b>	7.4
DAWL.7	<b>48.6</b>	7.4	<b>25.9</b>	3.5	<b>18.7</b>	2.9	<b>13.9</b>	2.9	<b>9.8</b>	6.4

	Description
DAWL.1 (eNet*)	Aug: $\hat{\beta}_{indv}$ — eNet using our alternative search
DAWL.2	Aug: $\hat{\beta}_{indv}$ , scaled
DAWL.3	Aug: $\hat{\beta}_{r-lasso}$ , scaled
DAWL.4	Aug: $\hat{\beta}_{r-lasso}$ , scaled; Wts: $\hat{\beta}_{r-lasso}$
DAWL.5	Aug: $\hat{\beta}_{r-lasso}$ , scaled; Wts: $\hat{\beta}_{ridge}$
DAWL.6	Aug: $\hat{\beta}_{r-lasso}$ , scaled; Wts: augmented $\hat{\beta}_{ols}$
DAWL.7	Aug: $\hat{\beta}_{r-lasso}$ , scaled; Wts: augmented $\hat{\beta}_{ols}$ 's $z$ -values (conditional)

possible, but limited. Here, we find that the data-dependent scaling (DAWL.2) can help a lot. On the other hand, a bad  $\hat{\beta}_{aug}$  indicates that a smaller value of  $\lambda$  is often preferred in (2.23), so the estimate would not be too far away from the lasso estimate and the sparsity can be maintained to some extent. If  $\hat{\beta}_{aug}$  is good but not sparse, as usually happens, (see DAWL.3-7), it is inevitable that we trade some sparsity for test error, since our first important criterion in this chapter is always test error and our search is designed to serve this.<sup>4</sup> Fortunately, adding weights can bring benefits to the sparsity.

To design a method combining the virtues of lasso and ridge is desirable. We see that DAWL offers a natural way to achieve this for  $\hat{\beta}_{lasso}$  and *any* given  $\hat{\beta}_{aug}$ . And it is interesting to notice that this fusion may yield a new estimate with even smaller test error than any of the two estimates.

## 2.4 Discussion

Motivated by our study of how ridge regression solves the singularity problem and how the lasso solves the sparsity problem, the data-augmentation technique was generalized and further developed. It is interesting to think of the data augmentation via  $\hat{\beta}_{aug}$  as an empirical Bayesian method with a simple Gaussian prior. The data-dependent scaling corresponds to a multivariate Gaussian with nondiagonal and degenerate covariance matrix. The nondiagonal data augmentation is more robust than the ordinary diagonal way since we are not using all the information provided by  $\hat{\beta}_{aug}$  in DA, and we do not have to, because of the tuning from  $\lambda$ . And as a future try, we may also include the standard error information of  $\hat{\beta}_{aug}$  into the model. We are also eager to learn in theory why the ranged lasso has such a small test-error. Compared to the success of data augmentation, the weight technique, although intuitive and having good asymptotics, showed somewhat limited power in reducing the test error. It is necessary to carry out more investigations of the weight construction in practice. Finally, it remains a problem to develop a scheme that can give explicit control over the trade-off between test error and sparsity.

---

<sup>4</sup>By the way, if test error is not our first concern, that is, we care more about sparsity than the test error, we can try a different  $\hat{\beta}_{aug}$  and/or redesign a regularization parameter search (like favoring small values of  $\lambda$ ) to achieve a new compromise between the two criteria.

## 2.5 Proof of Theorem 1

For any given  $\lambda_1, \lambda_2 > 0$ , we know that

$$\begin{aligned} & \left\| \begin{bmatrix} \mathbf{y} \\ \sqrt{\lambda_2} \mathbf{X}^T \mathbf{y} \end{bmatrix} - \begin{bmatrix} \mathbf{X} \\ \sqrt{\lambda_2} \mathbf{I} \end{bmatrix} \boldsymbol{\beta} \right\|_2^2 + \lambda_1 \|\boldsymbol{\beta}\|_1 \\ = & \boldsymbol{\beta}^T (\mathbf{X}^T \mathbf{X} + \lambda_2 \mathbf{I}) \boldsymbol{\beta} - 2(\lambda_2 + 1) \mathbf{y}^T \mathbf{X} \boldsymbol{\beta} + \lambda_1 \|\boldsymbol{\beta}\|_1 + (\mathbf{y}^T \mathbf{y} + \lambda_2 \mathbf{y}^T \mathbf{X} \mathbf{X}^T \mathbf{y}). \end{aligned}$$

So

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{\lambda_1, \lambda_2} &= \arg \min \boldsymbol{\beta}^T (\mathbf{X}^T \mathbf{X} + \lambda_2 \mathbf{I}) \boldsymbol{\beta} - 2(\lambda_2 + 1) \mathbf{y}^T \mathbf{X} \boldsymbol{\beta} + \lambda_1 \|\boldsymbol{\beta}\|_1 \\ &= \arg \min \boldsymbol{\beta}^T \left( \frac{\mathbf{X}^T \mathbf{X} + \lambda_2 \mathbf{I}}{1 + \lambda_2} \right) \boldsymbol{\beta} - 2 \mathbf{y}^T \mathbf{X} \boldsymbol{\beta} + \frac{\lambda_1}{1 + \lambda_2} \|\boldsymbol{\beta}\|_1. \end{aligned}$$

Comparing  $\hat{\boldsymbol{\beta}}_{\lambda_1, \lambda_2}$  to the definition of  $\hat{\boldsymbol{\beta}}_{\lambda_1, \lambda_2}^{(\epsilon Net)}$  yields the conclusion in Theorem 1.

## 2.6 Proof of Theorem 2

In the appendix, all inequalities and the absolute value ‘|’ are understood in the componentwise sense. First, we quote the following useful facts. Given a nonsingular matrix  $\mathbf{A} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix}$ , if  $\mathbf{A}_{11}$  is invertible, then  $\mathbf{S}_2 = \mathbf{A}_{22} - \mathbf{A}_{21} \mathbf{A}_{11}^{-1} \mathbf{A}_{12}$  is invertible, and

$$\mathbf{A}^{-1} = \begin{bmatrix} \mathbf{A}_{11}^{-1} + \mathbf{A}_{11}^{-1} \mathbf{A}_{12} \mathbf{S}_2^{-1} \mathbf{A}_{21} \mathbf{A}_{11}^{-1} & -\mathbf{A}_{11}^{-1} \mathbf{A}_{12} \mathbf{S}_2^{-1} \\ -\mathbf{S}_2^{-1} \mathbf{A}_{21} \mathbf{A}_{11}^{-1} & \mathbf{S}_2^{-1} \end{bmatrix}. \quad (2.24)$$

Similarly, if  $\mathbf{A}_{22}$  is invertible, then

$$\mathbf{A}^{-1} = \begin{bmatrix} \mathbf{S}_1^{-1} & -\mathbf{S}_1^{-1} \mathbf{A}_{12} \mathbf{A}_{22}^{-1} \\ -\mathbf{A}_{22}^{-1} \mathbf{A}_{21} \mathbf{S}_1^{-1} & \mathbf{A}_{22}^{-1} + \mathbf{A}_{22}^{-1} \mathbf{A}_{21} \mathbf{S}_1^{-1} \mathbf{A}_{12} \mathbf{A}_{22}^{-1} \end{bmatrix}, \quad (2.25)$$

where  $\mathbf{S}_1 = \mathbf{A}_{11} - \mathbf{A}_{12} \mathbf{A}_{22}^{-1} \mathbf{A}_{21}$ .

From (2.24) and (2.25), we further know that

$$\mathbf{A}^{-1} = \begin{bmatrix} \mathbf{A}_{11}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} + \begin{bmatrix} -\mathbf{A}_{11}^{-1}\mathbf{A}_{12} & \\ & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{S}_2^{-1} & \\ & \mathbf{S}_2^{-1} \end{bmatrix} \begin{bmatrix} -\mathbf{A}_{21}\mathbf{A}_{11}^{-1} & \mathbf{I} \\ -\mathbf{A}_{21}\mathbf{A}_{11}^{-1} & \mathbf{I} \end{bmatrix} \quad (2.26)$$

$$\mathbf{A}^{-1} = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_{22}^{-1} \end{bmatrix} + \begin{bmatrix} \mathbf{I} & \\ & -\mathbf{A}_{22}^{-1}\mathbf{A}_{21} \end{bmatrix} \begin{bmatrix} \mathbf{S}_1^{-1} & \\ & \mathbf{S}_1^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{I} & -\mathbf{A}_{12}\mathbf{A}_{22}^{-1} \\ \mathbf{I} & -\mathbf{A}_{12}\mathbf{A}_{22}^{-1} \end{bmatrix} \quad (2.27)$$

Now go back to the lasso estimate  $\hat{\boldsymbol{\beta}}$  in the settings of Theorem 2. It is known, say, from [42] that  $\hat{\boldsymbol{\beta}}$  is an optimal solution if and only if there exists a subgradient  $\mathbf{s} \in \partial\|\boldsymbol{\beta}\|_1$  such that

$$\boldsymbol{\Sigma}\hat{\boldsymbol{\beta}} = \mathbf{X}^T\mathbf{y} - \lambda\mathbf{s},$$

where  $\mathbf{s} = [s_i]$  satisfies  $s_i = 1, -1$ , or  $\in [-1, 1]$ , if  $\beta_i > 0, < 0$ , or  $= 0$ . We can write it as

$$\boldsymbol{\Sigma}\hat{\boldsymbol{\beta}} = \mathbf{X}^T\mathbf{y} - \lambda\widetilde{\text{sgn}}(\hat{\boldsymbol{\beta}}),$$

or

$$\boldsymbol{\Sigma}\hat{\boldsymbol{\beta}} = \begin{bmatrix} \boldsymbol{\Sigma}_{12}\boldsymbol{\beta}_2 + \mathbf{X}_1^T\boldsymbol{\epsilon} - \lambda\widetilde{\text{sgn}}(\hat{\boldsymbol{\beta}}_1) \\ \boldsymbol{\Sigma}_2\boldsymbol{\beta}_2 + \mathbf{X}_2^T\boldsymbol{\epsilon} - \lambda\widetilde{\text{sgn}}(\hat{\boldsymbol{\beta}}_2) \end{bmatrix} \quad (2.28)$$

where  $\widetilde{\text{sgn}}(\hat{\boldsymbol{\beta}}) = [\widetilde{\text{sgn}}(\hat{\beta}_i)]_{d \times 1}$  is a simple notation (not a function); for the lasso estimate  $\hat{\boldsymbol{\beta}}$ ,  $\widetilde{\text{sgn}}(\hat{\beta}_i)$  defined by (2.28) takes 1,  $-1$ , or some value in  $[-1, 1]$  (unique) when  $\hat{\beta}_i > 0, < 0$ , or  $= 0$ , respectively.

Assume  $\boldsymbol{\Sigma}$  is nonsingular. It follows from (2.27) that

**Lemma 1** *The lasso estimate can be represented as*

$$\hat{\boldsymbol{\beta}}_1 = \boldsymbol{\Sigma}_1^{-1}(\mathbf{X}_1^T\boldsymbol{\epsilon} - \lambda\widetilde{\text{sgn}}(\hat{\boldsymbol{\beta}}_1)) - \boldsymbol{\Sigma}_1^{-1}\boldsymbol{\Sigma}_{12}(\hat{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_2) \quad (2.29)$$

$$\hat{\boldsymbol{\beta}}_2 = \boldsymbol{\beta}_2 + \mathbf{S}_2^{-1} \left[ (\mathbf{X}_2^T - \boldsymbol{\Sigma}_{12}^T\boldsymbol{\Sigma}_1^{-1}\mathbf{X}_1^T)\boldsymbol{\epsilon} + \lambda\boldsymbol{\Sigma}_{12}^T\boldsymbol{\Sigma}_1^{-1}\widetilde{\text{sgn}}(\hat{\boldsymbol{\beta}}_1) - \lambda\widetilde{\text{sgn}}(\hat{\boldsymbol{\beta}}_2) \right] \quad (2.30)$$

or,

$$\mathbf{S}_1\hat{\boldsymbol{\beta}}_1 = \left[ (\mathbf{X}_1^T - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_2^{-1}\mathbf{X}_2^T)\boldsymbol{\epsilon} + \lambda\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_2^{-1}\widetilde{\text{sgn}}(\hat{\boldsymbol{\beta}}_2) - \lambda\widetilde{\text{sgn}}(\hat{\boldsymbol{\beta}}_1) \right] \quad (2.31)$$

$$\hat{\boldsymbol{\beta}}_2 = \boldsymbol{\beta}_2 + \boldsymbol{\Sigma}_2^{-1}(\mathbf{X}_2^T\boldsymbol{\epsilon} - \lambda\widetilde{\text{sgn}}(\hat{\boldsymbol{\beta}}_2)) - \boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_{12}^T\hat{\boldsymbol{\beta}}_1. \quad (2.32)$$

(2.29)–(2.32) are our basic starting point.

**Lemma 2** *Let  $\mathbf{z} \sim N(\mathbf{0}, \mathbf{D}_{d \times d})$ ,  $\mathbf{z}' \sim N(\mathbf{0}, \mathbf{\Lambda}_{d \times d})$ , where  $\mathbf{D}$  is a diagonal matrix, and the diagonal entries of  $\mathbf{\Lambda}$ , denoted by  $\text{diag}(\mathbf{\Lambda})$ , are the same as those of  $\mathbf{D}$ , i.e.,  $\text{diag}(\mathbf{\Lambda}) = \text{diag}(\mathbf{D})$ . Then*

$$P(\max |\mathbf{z}'| \geq \lambda) \leq P(\max |\mathbf{z}| \geq \lambda), \quad (2.33)$$

for any  $\lambda$ .

This is clear from Šidák's classical result [52] in 1967.

1. Sign consistency. Let  $A = \{\hat{\boldsymbol{\beta}}_1 = \mathbf{0}\}$ . We estimate  $P(A^c)$ . First, by KKT and (2.31),

$$\begin{aligned} \hat{\boldsymbol{\beta}}_1 = \mathbf{0} &\iff \hat{\boldsymbol{\beta}}_1 = \mathbf{0} \text{ and } |\widehat{\text{sgn}}(\hat{\boldsymbol{\beta}}_1)| \leq 1 \\ &\iff \left| (\mathbf{X}_1^T - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_2^{-1}\mathbf{X}_2^T)\boldsymbol{\epsilon} + \lambda\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_2^{-1}\widehat{\text{sgn}}(\hat{\boldsymbol{\beta}}_2) \right| \leq \lambda. \end{aligned} \quad (2.34)$$

A more careful look shows the assumption of the nonsingularity of  $\boldsymbol{\Sigma}_2$ , but not  $\boldsymbol{\Sigma}$ , is enough in the second equivalence. Let  $\boldsymbol{\Sigma}_{12} = [\mathbf{v}_1, \dots, \mathbf{v}_{d_1}]^T$ , then  $\|\mathbf{v}_i\|_2 \leq \kappa\sqrt{d_2}$ . So  $|\mathbf{v}_i^T \boldsymbol{\Sigma}_2^{-1} \widehat{\text{sgn}}(\hat{\boldsymbol{\beta}}_2)| \leq \|\mathbf{v}_i\|_2 \cdot \|\boldsymbol{\Sigma}_2^{-1}\|_2 \cdot \|\widehat{\text{sgn}}(\hat{\boldsymbol{\beta}}_2)\|_2 \leq \kappa d_2 / \tau_2$ . And  $|\boldsymbol{\Sigma}_{12} \boldsymbol{\Sigma}_2^{-1} \widehat{\text{sgn}}(\hat{\boldsymbol{\beta}}_2)| \leq \kappa d_2 / \tau_2$ . Let  $\mathbf{X}_1'^T = \mathbf{X}_1^T - \boldsymbol{\Sigma}_{12} \boldsymbol{\Sigma}_2^{-1} \mathbf{X}_2^T$ . A sufficient condition of (2.34) is

$$\max \left| \mathbf{X}_1'^T \boldsymbol{\epsilon} \right| + \lambda \cdot \kappa d_2 / \tau_2 \leq \lambda,$$

or

$$\max \left| \mathbf{X}_1'^T \boldsymbol{\epsilon} \right| \leq (1 - \kappa d_2 / \tau_2) \lambda. \quad (2.35)$$

Define  $B = \left\{ \max \left| \mathbf{X}_1'^T \boldsymbol{\epsilon} \right| \leq (1 - \kappa d_2 / \tau_2) \lambda \right\}$ . Then  $P(A^c) \leq P(B^c)$ .

Let  $\boldsymbol{\epsilon}'_1 = \mathbf{X}_1'^T \boldsymbol{\epsilon} \in \Re^{d_1}$ . Note that  $\mathbf{X}_1'^T \mathbf{X}_1' = \mathbf{S}_1$ , so  $\boldsymbol{\epsilon}'_1 \sim N(\mathbf{0}, \sigma^2 \mathbf{S}_1)$ . Since  $\text{diag}(\mathbf{S}_1) = \mathbf{1}$  and  $\text{diag}(\boldsymbol{\Sigma}_{12} \boldsymbol{\Sigma}_2^{-1} \boldsymbol{\Sigma}_{12}^T) = [\mathbf{v}_i^T \boldsymbol{\Sigma}_2^{-1} \mathbf{v}_i] \geq \mathbf{0}$ ,  $\text{diag}(\mathbf{S}_1) \leq \mathbf{1}$ . Lemma 2 tells us that

$$P(B^c) \leq P \left( \max |\boldsymbol{\epsilon}''_1| \cdot \sigma \geq \lambda \left( 1 - \frac{\kappa d_2}{\tau_2} \right) \right),$$

where  $\boldsymbol{\epsilon}_1'' \sim N(\mathbf{0}, \mathbf{I}_{d_1 \times d_1})$ . So

$$P(B^c) \leq d_1 P\left(|\epsilon_{1,i}''| \geq \lambda\left(1 - \frac{\kappa d_2}{\tau_2}\right) \frac{1}{\sigma}\right) \equiv 2d_1 \Phi([M, +\infty)).$$

Using the standard bound of the normal tail probability, we get

$$P(A^c) \leq P(B^c) \leq 2d_1 \frac{1}{M} \varphi(M). \quad (2.36)$$

Furthermore, we know that (this also depends on the nonsingularity of  $\boldsymbol{\Sigma}_2$  only)

$$\begin{aligned} \hat{\boldsymbol{\beta}}_1 = \mathbf{0}, \hat{\beta}_{2,i} \neq 0, \forall i : 1 \leq i \leq d_2 &\iff \left| \mathbf{X}_1'^T \boldsymbol{\epsilon} + \lambda \boldsymbol{\Sigma}_{12} \boldsymbol{\Sigma}_2^{-1} \widetilde{\text{sgn}}(\hat{\boldsymbol{\beta}}_2) \right| \leq \lambda, \text{ and} \\ \hat{\boldsymbol{\beta}}_2 = \boldsymbol{\beta}_2 + \boldsymbol{\Sigma}_2^{-1}(\mathbf{X}_2^T \boldsymbol{\epsilon} - \lambda \text{sgn}(\hat{\boldsymbol{\beta}}_2)), \text{sgn}(\hat{\beta}_{2,i}) \neq 0, \forall i : 1 \leq i \leq d_2. & \end{aligned} \quad (2.37)$$

Since  $\left| \boldsymbol{\beta}_2 + \boldsymbol{\Sigma}_2^{-1}(\mathbf{X}_2^T \boldsymbol{\epsilon} - \lambda \text{sgn}(\hat{\boldsymbol{\beta}}_2)) \right| \geq |\boldsymbol{\beta}_2| - \lambda \left| \boldsymbol{\Sigma}_2^{-1} \text{sgn}(\hat{\boldsymbol{\beta}}_2) \right| - \left| \boldsymbol{\Sigma}_2^{-1} \mathbf{X}_2^T \boldsymbol{\epsilon} \right|$ , a sufficient condition of (2.37) is:

$$\begin{aligned} \left| \mathbf{X}_1' \boldsymbol{\epsilon} + \lambda \boldsymbol{\Sigma}_{12} \boldsymbol{\Sigma}_2^{-1} \widetilde{\text{sgn}}(\hat{\boldsymbol{\beta}}_2) \right| &\leq \lambda, \text{ and} \\ \left| \boldsymbol{\Sigma}_2^{-1} \mathbf{X}_2^T \boldsymbol{\epsilon} \right| + \lambda \left| \boldsymbol{\Sigma}_2^{-1} \mathbf{s} \right| &\leq |\boldsymbol{\beta}_2| \text{ for any } \mathbf{s} \text{ satisfying } |\mathbf{s}| \leq 1. \end{aligned}$$

In the situation of  $\beta_{2,i} \neq 0$  for all  $i$ , this is still sufficient for  $\hat{\boldsymbol{\beta}}$  to have  $\text{sgn}(\hat{\boldsymbol{\beta}}) = \text{sgn}(\boldsymbol{\beta})$ .

Using the spectral decomposition of  $\boldsymbol{\Sigma}_2$ :  $\boldsymbol{\Sigma}_2 = \mathbf{U} \mathbf{D} \mathbf{U}^T$  with  $\mathbf{U} = [\mathbf{u}_1, \dots, \mathbf{u}_{d_2}]^T$ , we can represent  $\boldsymbol{\Sigma}_2^{-1}$  as  $[\mathbf{u}_i^T \mathbf{D}^{-1} \mathbf{u}_j]_{d_2 \times d_2}$ , and  $\boldsymbol{\Sigma}_2^{-1} \mathbf{s}$  as  $[\sum_{j=1}^{d_2} s_j \mathbf{u}_i^T \mathbf{D}^{-1} \mathbf{u}_j]_{d_2 \times 1}$ , which means that  $\text{diag}(\boldsymbol{\Sigma}_2^{-1}) \leq \frac{1}{\tau_2}$ ,  $|\boldsymbol{\Sigma}_2^{-1} \mathbf{s}| \leq \frac{d_2}{\tau_2}$ . And a simpler sufficient condition of (2.37) can be obtained

$$\left| \mathbf{X}_1'^T \boldsymbol{\epsilon} \right| \leq (1 - \kappa d_2 / \tau_2) \lambda \text{ and } \left| \boldsymbol{\Sigma}_2^{-1} \mathbf{X}_2^T \boldsymbol{\epsilon} \right| \leq L_0, \quad (2.38)$$

where  $L_0 = \min |\boldsymbol{\beta}_2| - \lambda d_2 / \tau_2$ . Denote by  $V$  the set of  $\boldsymbol{\epsilon}$  satisfying the second condition.

Note that  $\boldsymbol{\Sigma}_2^{-1} \mathbf{X}_2^T \boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 \boldsymbol{\Sigma}_2^{-1})$ , and we've shown that  $\text{diag}(\boldsymbol{\Sigma}_2^{-1}) \leq \frac{1}{\tau_2}$ . By Lemma 2 again,

$$P(V^c) = P(\max |\boldsymbol{\Sigma}_2^{-1} \mathbf{X}_2^T \boldsymbol{\epsilon}| \geq L_0) \leq P(\max |\epsilon_2''| \geq L_0 \frac{\sqrt{\tau_2}}{\sigma}) \leq 2d_2 \Phi([L, \infty)),$$

where  $\boldsymbol{\epsilon}_2'' \sim N(\mathbf{0}, \mathbf{I}_{d_2 \times d_2})$ . So

$$\begin{aligned} P\left(\text{sgn}(\hat{\boldsymbol{\beta}}) \neq \text{sgn}(\boldsymbol{\beta})\right) &\leq P(A^c \cup V^c) \leq P(A^c) + P(V^c) \\ &\leq 2d_1 \frac{1}{M} \varphi(M) + 2d_2 \frac{1}{L} \varphi(L). \end{aligned} \quad (2.39)$$

In fact, we can get something slightly stronger than (2.39). Observing that  $\mathbf{X}_1'^T \boldsymbol{\epsilon}$  is independent of  $\mathbf{X}_2^T \boldsymbol{\epsilon}$ , we have

$$\begin{aligned} P\left(\text{sgn}(\hat{\boldsymbol{\beta}}) = \text{sgn}(\boldsymbol{\beta})\right) &\geq P(A \cap V) = P(A) \cdot P(V) \\ &\geq [1 - 2\Phi(-M)]^{d_1} [1 - 2\Phi(-L)]^{d_2}. \end{aligned} \quad (2.40)$$

(2.40) implies (2.39).

2. Risks. Next we deal with the risks  $R_1, R_2$ . Let  $r_1 = \|\hat{\boldsymbol{\beta}}_1\|_2^2$ ,  $r_2 = \|\hat{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_2\|_2^2$ . In the following part, we list several ways to estimate the risks roughly.

From (2.32),

$$\begin{aligned} r_2 &= \|\boldsymbol{\Sigma}_2^{-1}(\mathbf{X}_2^T \boldsymbol{\epsilon} - \lambda \widetilde{\text{sgn}}(\hat{\boldsymbol{\beta}}_2)) - \boldsymbol{\Sigma}_2^{-1} \boldsymbol{\Sigma}_{12}^T \hat{\boldsymbol{\beta}}_1\|_2^2 \implies \\ r_2 &\leq c_2 \|\boldsymbol{\Sigma}_2^{-1}(\mathbf{X}_2^T \boldsymbol{\epsilon} - \lambda \widetilde{\text{sgn}}(\hat{\boldsymbol{\beta}}_2))\|_2^2 + c_2 \cdot \kappa^2 \frac{d_1 d_2}{\tau_2^2} \cdot r_1, \end{aligned} \quad (2.41)$$

by Cauchy-Schwarz inequality with  $c_2 = 2$  and the fact that  $\|\boldsymbol{\Sigma}_{12}\|_2 = \max_{\|\boldsymbol{\alpha}\|_2=1} \|\boldsymbol{\Sigma}_{12} \boldsymbol{\alpha}\|_2 \leq \kappa \sqrt{d_1 d_2}$ . From (2.29),

$$\begin{aligned} r_1 &= \|\boldsymbol{\Sigma}_1^{-1}(\mathbf{X}_1^T \boldsymbol{\epsilon} - \lambda \widetilde{\text{sgn}}(\hat{\boldsymbol{\beta}}_1)) - \boldsymbol{\Sigma}_1^{-1} \boldsymbol{\Sigma}_{12}(\hat{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_2)\|_2^2 \implies \\ r_1 &\leq c_1 \|\boldsymbol{\Sigma}_1^{-1}(\mathbf{X}_1^T \boldsymbol{\epsilon} - \lambda \widetilde{\text{sgn}}(\hat{\boldsymbol{\beta}}_1))\|_2^2 + c_1 \kappa^2 \frac{d_1 d_2}{\tau_1^2} \cdot r_2. \end{aligned}$$

To take advantage of the thresholding effect, we write it as

$$r_1 \leq c_1 \|\boldsymbol{\Sigma}_1^{-1}(\mathbf{X}_1^T \boldsymbol{\epsilon} - \lambda \widetilde{\text{sgn}}(\hat{\boldsymbol{\beta}}_1))\|_2^2 \cdot 1_{A^c} + c_1 \cdot \kappa^2 \frac{d_1 d_2}{\tau_1^2} \cdot r_2 1_{A^c}, \quad (2.42)$$

where  $c_1 = 2$ .

Inserting (2.42) into (2.41) we get

$$\begin{aligned}
 r_2 &\leq c_2 \|\Sigma_2^{-1}(\mathbf{X}_2^T \boldsymbol{\epsilon} - \lambda \widetilde{\text{sgn}}(\hat{\boldsymbol{\beta}}_2))\|_2^2 \\
 &\quad + c_2 \cdot \kappa^2 \frac{d_1 d_2}{\tau_2^2} \cdot c_1 \|\Sigma_1^{-1}(\mathbf{X}_1^T \boldsymbol{\epsilon} - \lambda \widetilde{\text{sgn}}(\hat{\boldsymbol{\beta}}_1))\|_2^2 \cdot 1_{A^c} \\
 &\quad + c_2 \cdot \kappa^2 \frac{d_1 d_2}{\tau_2^2} \cdot c_1 \cdot \kappa^2 \frac{d_1 d_2}{\tau_1^2} \cdot r_2 1_{A^c},
 \end{aligned}$$

and thus

$$\begin{aligned}
 \left(1 - c_1 c_2 \cdot \kappa^4 \frac{d_1^2 d_2^2}{\tau_1^2 \tau_2^2}\right) r_2 &\leq c_2 \|\Sigma_2^{-1}(\mathbf{X}_2^T \boldsymbol{\epsilon} - \lambda \widetilde{\text{sgn}}(\hat{\boldsymbol{\beta}}_2))\|_2^2 \\
 &\quad + c_1 c_2 \cdot \kappa^2 \frac{d_1 d_2}{\tau_2^2} \|\Sigma_1^{-1}(\mathbf{X}_1^T \boldsymbol{\epsilon} - \lambda \widetilde{\text{sgn}}(\hat{\boldsymbol{\beta}}_1))\|_2^2 \cdot 1_{A^c}.
 \end{aligned} \tag{2.43}$$

Similarly, we can get

$$\begin{aligned}
 \left(1 - c_1 c_2 \cdot \kappa^4 \frac{d_1^2 d_2^2}{\tau_1^2 \tau_2^2}\right) r_1 &\leq c_1 \|\Sigma_1^{-1}(\mathbf{X}_1^T \boldsymbol{\epsilon} - \lambda \widetilde{\text{sgn}}(\hat{\boldsymbol{\beta}}_1))\|_2^2 \cdot 1_{A^c} \\
 &\quad + c_1 c_2 \cdot \kappa^2 \frac{d_1 d_2}{\tau_1^2} \|\Sigma_2^{-1}(\mathbf{X}_2^T \boldsymbol{\epsilon} - \lambda \widetilde{\text{sgn}}(\hat{\boldsymbol{\beta}}_2))\|_2^2 \cdot 1_{A^c}.
 \end{aligned} \tag{2.44}$$

The formulas (2.29)–(2.32), (2.41)–(2.44) can be used to estimate the risks of  $\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2$ . For instance, the results in Theorem 2 are obtained from (2.41), and (2.31). By the way, the factors  $c_1, c_2$  for  $R_1$  and  $R_2$  rising from applying Cauchy-Schwarz inequalities may be recomputed more carefully regarding the total number of terms in the end.

Taking expectation on both sides of (2.41) (with 3 terms), we see

$$R_2 \leq 3 \left( E \|\Sigma_2^{-1} \mathbf{X}_2^T \boldsymbol{\epsilon}\|_2^2 + \lambda^2 E \|\Sigma_2^{-1} \widetilde{\text{sgn}}(\hat{\boldsymbol{\beta}}_2)\|_2^2 + \kappa^2 \frac{d_1 d_2}{\tau_2^2} \cdot R_1 \right),$$

and thus

$$R_2 \leq 3 \left( \sigma^2 \cdot \text{tr}(\Sigma_2^{-1}) + \lambda^2 \frac{d_2}{\tau_2^2} + \kappa^2 \frac{d_1 d_2}{\tau_2^2} \cdot R_1 \right). \tag{2.45}$$

Hence (2.7) holds.

To get (2.8), we need the following result about  $\lambda_{\max}(\mathbf{S}_1^{-1})$ , the largest eigenvalue of

$\mathbf{S}_1^{-1}$ :

$$\lambda_{max}(\mathbf{S}_1^{-1}) \leq \frac{1}{\tau_1} \left( 1 - \kappa^2 \cdot \frac{d_1 d_2}{\tau_1 \tau_2} \right)^{-1}. \quad (2.46)$$

This is true by noting that  $\mathbf{S}_1 = \mathbf{X}'_1{}^T \mathbf{X}'_1$  is semi-positive definite and  $\lambda_{min}(\mathbf{S}_1) \geq \tau_1 - \kappa^2 d_1 d_2 / \tau_2$ .

By (2.31), we have

$$\begin{aligned} r_1 &\leq 3\lambda_{max}^2(\mathbf{S}_1^{-1}) \left[ \|\mathbf{X}'_1{}^T \boldsymbol{\epsilon}\|_2^2 + \lambda^2 \|\boldsymbol{\Sigma}_{12} \boldsymbol{\Sigma}_2^{-1} \widehat{\text{sgn}}(\hat{\boldsymbol{\beta}}_2)\|_2^2 + \lambda^2 \|\widehat{\text{sgn}}(\hat{\boldsymbol{\beta}}_1)\|_2^2 \right] \cdot 1_{A^c} \\ &\leq 3\lambda_{max}^2(\mathbf{S}_1^{-1}) \left[ \|\mathbf{X}'_1{}^T \boldsymbol{\epsilon}\|_2^2 \cdot 1_{A^c} + \left( \lambda^2 d_1 \cdot \kappa^2 \frac{d_2^2}{\tau_2^2} + \lambda^2 d_1 \right) \cdot 1_{A^c} \right]. \end{aligned}$$

Thus

$$\begin{aligned} R_1 &\leq 3\lambda_{max}^2(\mathbf{S}_1^{-1}) \left[ E(\|\mathbf{X}'_1{}^T \boldsymbol{\epsilon}\|_2^2 \cdot 1_{A^c}) + \left( 1 + \kappa^2 \frac{d_2^2}{\tau_2^2} \right) \lambda^2 d_1 \cdot 2d_1 \frac{1}{M} \varphi(M) \right] \\ &\leq 3\lambda_{max}^2(\mathbf{S}_1^{-1}) \left[ E(\|\boldsymbol{\epsilon}'_1\|_2^2 \cdot 1_{B^c}) + \left( 1 + \kappa^2 \frac{d_2^2}{\tau_2^2} \right) 2\lambda^2 d_1^2 \frac{1}{M} \varphi(M) \right] \\ &\leq 3\lambda_{max}^2(\mathbf{S}_1^{-1}) \cdot d_1 E \left[ (\max |\boldsymbol{\epsilon}'_1|)^2; \max |\boldsymbol{\epsilon}'_1| \geq \lambda \left( 1 - \kappa \frac{d_2}{\tau_2} \right) \right] + \\ &\quad 3\lambda_{max}^2(\mathbf{S}_1^{-1}) \cdot \frac{1 + \kappa^2 d_2^2 / \tau_2^2}{(1 - \kappa d_2 / \tau_2)^2} \cdot 2\sigma^2 d_1^2 M \varphi(M). \end{aligned}$$

Since for a random variable  $z$  with probability density  $p(\cdot)$  and  $a > 0$ ,

$$E(z^2; z \geq a) = \int_a^\infty t^2 p(t) dt = \int_a^\infty 2sP(z \geq s) ds + a^2 P(z \geq a),$$

it follows from Lemma 2 that

$$E \left[ (\max |\boldsymbol{\epsilon}'_1|)^2; \max |\boldsymbol{\epsilon}'_1| \geq \lambda \left( 1 - \kappa \frac{d_2}{\tau_2} \right) \right] \leq E \left[ \max |\boldsymbol{\epsilon}''_1|^2 \cdot \sigma^2; \max |\boldsymbol{\epsilon}''_1| \geq M \right].$$

(Recall that  $\boldsymbol{\epsilon}''_1 \sim N(\mathbf{0}, \mathbf{I}_{d_1 \times d_1})$ .) The density of  $\max |\boldsymbol{\epsilon}''_1|$  is given by  $2d_1 \varphi(t) (1 - 2\Phi(-t))^{d_1 - 1}$ .

It is easy to get

$$E \left[ \max |\boldsymbol{\epsilon}''_1|^2; \max |\boldsymbol{\epsilon}''_1| \geq M \right] \leq 2d_1 \int_M^\infty t^2 \varphi(t) dt \leq 2d_1 (M + 1/M) \varphi(M).$$

Hence,

$$\begin{aligned}
 R_1 &\leq 3\frac{1}{\tau_1^2} \left(1 - \kappa^2 \frac{d_1 d_2}{\tau_1 \tau_2}\right)^{-2} \cdot \left(2\sigma^2 d_1^2 \left(M + \frac{1}{M}\right) \varphi(M) \right. \\
 &\quad \left. + \frac{1 + \kappa^2 d_2^2 / \tau_2^2}{(1 - \kappa d_2 / \tau_2)^2} \cdot 2\sigma^2 d_1^2 M \varphi(M)\right) \\
 &= \frac{\sigma^2}{\tau_1^2} d_1^2 (K_1 M + K_2 \frac{1}{M}) \varphi(M).
 \end{aligned}$$

The proof is now complete.

We may use other formulas of (2.29)–(2.32), (2.41)–(2.44), too, in estimating  $R_1, R_2$  (notice the fact that  $\mathbf{X}'_1 \boldsymbol{\epsilon}$  is independent of  $\mathbf{X}'_2 \boldsymbol{\epsilon}$ ). Finally, bounds for  $E\|\hat{\boldsymbol{\beta}}_1\|$  and  $E\|\hat{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_2\|$  can be similarly obtained.

## 2.7 The Ranged Lasso Process

Assume  $\mathbf{X}$  is orthogonal. Let  $\hat{\boldsymbol{\beta}}$  be the solution of

$$\min \|\boldsymbol{\beta} - \boldsymbol{\beta}_{obs}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1 \quad \text{s.t.} \quad \max \boldsymbol{\beta} - \min \boldsymbol{\beta} \leq C,$$

where  $\lambda \geq 0$ . It is equivalent to

$$\min \|\boldsymbol{\beta} - \boldsymbol{\beta}_{obs}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1 \quad \text{s.t.} \quad u - l \leq C, l \leq \beta_i \leq u, \forall i : 1 \leq i \leq d, .$$

Define the Lagrangian to be

$$L = \|\boldsymbol{\beta} - \boldsymbol{\beta}_{obs}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1 + \gamma(u - l - C) + \sum \eta_i(\beta_i - u) + \sum \tau_i(l - \beta_i).$$

The KKT conditions can be obtained:

$$\left\{ \begin{array}{l} \hat{\beta} = \beta_{obs} - \frac{\lambda}{2} \widetilde{\text{sgn}}(\hat{\beta}) - \frac{1}{2} \eta + \frac{1}{2} \tau \\ \sum \eta_i = \sum \tau_i = \gamma \\ \gamma(u - l - C) = 0 \\ \eta_i(\hat{\beta}_i - u) = 0, \forall i \\ \tau_i(l - \hat{\beta}_i) = 0, \forall i \\ l \leq \hat{\beta}_i \leq u, \forall i \\ u - l \leq C \\ \eta, \tau, \lambda, \gamma \geq 0 \end{array} \right. \quad \text{or} \quad \left\{ \begin{array}{l} \hat{\beta} = \beta_{obs} - \frac{\lambda}{2} \widetilde{\text{sgn}}(\hat{\beta}) - \frac{1}{2} \eta + \frac{1}{2} \tau \\ \sum \eta_i = \sum \tau_i = \gamma \\ \gamma(\max \hat{\beta} - \min \hat{\beta} - C) = 0 \\ \eta_i(\hat{\beta}_i - \max \hat{\beta}) = 0, \forall i \\ \tau_i(\min \hat{\beta} - \hat{\beta}_i) = 0, \forall i \\ \max \hat{\beta} - \min \hat{\beta} \leq C \\ \eta, \tau, \lambda, \gamma \geq 0 \end{array} \right. .$$

Assume the range constraint is nontrivial. That is,  $C < \max \hat{\beta}_{lasso} - \min \hat{\beta}_{lasso}$ . So  $\gamma > 0$ ,  $\max \hat{\beta} - \min \hat{\beta} = C$ . Suppose  $\min \hat{\beta} \leq 0 \leq \max \hat{\beta}$ , and define  $H, K$  to be the index sets such that

$$\begin{aligned} \hat{\beta}_j &= \max \hat{\beta} \triangleq b, \forall j \in H \\ \hat{\beta}_j &= \min \hat{\beta} \triangleq a, \forall j \in K. \end{aligned}$$

Let  $(H \cup K)^c = N \cup O \cup P$ , with  $\hat{\beta}_j < 0, = 0$ , or  $> 0$  if  $j \in N, O$ , or  $P$ , respectively. Now the KKT conditions become

$$\left\{ \begin{array}{l} \hat{\beta} = \beta_{obs} - \frac{\lambda}{2} \widetilde{\text{sgn}}(\hat{\beta}) - \frac{1}{2} \eta + \frac{1}{2} \tau \\ \sum_{j \in K} \tau_j = \sum_{j \in H} \eta_j \\ \eta_j = 0, \forall j \notin H \\ \tau_j = 0, \forall j \notin K \\ b - a = C \\ \eta, \tau, \lambda \geq 0 \end{array} \right.$$

which are equivalent to

$$\left\{ \begin{array}{l} \hat{\beta}_j = \begin{cases} a (= \beta_{obs,j} + \frac{\lambda}{2} + \frac{1}{2}\tau_j), & \text{if } j \in K \\ \beta_{obs,j} + \frac{\lambda}{2}, & \text{if } j \in N \\ 0, & \text{if } j \in O \\ \beta_{obs,j} - \frac{\lambda}{2}, & \text{if } j \in P \\ b (= \beta_{obs,j} - \frac{\lambda}{2} - \frac{1}{2}\eta_j), & \text{if } j \in H \end{cases} \\ \sum_{j \in K} \tau_j = \sum_{j \in H} \eta_j \\ b - a = C \\ \boldsymbol{\eta}, \boldsymbol{\tau}, \lambda \geq 0 \end{array} \right.$$

Hence in the orthogonal case, evaluating  $\hat{\boldsymbol{\beta}}_{r-lasso}$  is a combined process of the lasso shrinkage and  $\max \boldsymbol{\beta} / \min \boldsymbol{\beta}$  auto-clustering. This is described in Section 2.1.4, with

$$\Delta_1 = \sum_{j \in K} \tau_j, \Delta_2 = \sum_{j \in H} \eta_j,$$

denoting the total movements toward  $a, b$  in Figure 2.1b.

In the non-orthogonal case, the clustering and the thresholding are still present, but can not be separated in such a clear way. We can view the criterion of (2.16) in the following way as an approximation to solve the corresponding  $l_0$  problem:

$$\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1 + \alpha_1 \sum_i |\beta_i - \min \boldsymbol{\beta}| + \alpha_2 \sum_i |\beta_i - \max \boldsymbol{\beta}|$$

under  $\alpha_1 = \alpha_2$ , which is in fact  $\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1 + \alpha(\max \boldsymbol{\beta} - \min \boldsymbol{\beta})$ .

## 2.8 Proof of Theorem 3

Let  $\hat{\beta}^H, \hat{\beta}^S$  denote the hard- and soft-thresholding estimates with threshold value  $\lambda_0\sigma$ . It is easy to see  $\hat{\beta}^H, \hat{\beta}^S$ , and  $\hat{\beta}$  all have the same sign and  $\hat{\beta}$  is sandwiched by the other two. Therefore,

$$E\|\hat{\beta} - \beta\|_2^2 \leq \sum E(\max((\hat{\beta}_i^S - \beta_i)^2, (\hat{\beta}_i^H - \beta_i)^2)).$$

We need to study soft- and hard-thresholdings in the univariate case.

Let  $y = \mu + \epsilon$  (all are scalars) with  $\epsilon \sim N(0, 1)$ , and  $\rho_S(\lambda, \mu), \rho_H(\lambda, \mu)$  be the risks of the soft- and hard-thresholdings with parameter  $\lambda$ . It is well known [20, 14] that

$$\rho_S(\lambda, \mu) \leq \min(\rho_S(\lambda, 0) + \mu^2, 1 + \lambda^2) \leq \min\left(\frac{2\varphi(\lambda)}{\lambda} + \mu^2, 1 + \lambda^2\right) \quad (2.47)$$

for any  $\lambda > 0$ . Yet we feel that there is *no* such explicit nonasymptotic bound, or a complete proof for the hard-thresholding rule. This short appendix is mainly to give some details about this.

Our goal is to show the following on the basis of [20]

$$\rho_H(\lambda, \mu) \leq 1 + \lambda^2 \text{ for } \lambda > 1 \quad (2.48)$$

$$\rho_H(\lambda, \mu) \leq \rho_H(\lambda, 0) + 1.2\mu^2. \quad (2.49)$$

Dohoho and Johnstone have shown (2.48), and (2.49) for  $0 < \mu < \lambda$ , but it is technically difficult to use the second derivative to prove (2.49) for any  $\mu > 0$ .

Let  $g = \partial\rho_H/\partial\mu - 2.4\mu$ , and  $\rho_H(\lambda, \mu)$  is known to be [20, 30]

$$1 + (\mu^2 - 1)(\Phi(\lambda - \mu) - \Phi(-\lambda - \mu)) + (\lambda + \mu)\varphi(\lambda + \mu) + (\lambda - \mu)\varphi(\lambda - \mu),$$

where  $\varphi, \Phi$  are the standard normal density and distribution functions, respectively. One may observe that

$$\sup_{\mu \geq 0} g(0, \mu) \leq \sup_{\lambda \geq 0} g(\lambda, 0) = 0,$$

which is trivial to verify. So it is sufficient to show that for any  $(\lambda, \mu) > 0$ , there exists some  $\theta \in [\pi, \frac{3}{2}\pi]$  such that the directional derivative  $D_\theta g$  at  $(\lambda, \mu)$  is greater than 0, or  $\exists \theta_{\lambda, \mu} \in [0, \frac{\pi}{2}]$  s.t.  $D_\theta g(\lambda, \mu) < 0$ , because  $g$  is smooth enough.

Consider a uniform direction  $\theta = \frac{\pi}{4}$ , and let  $h = D_{\theta}g = (\frac{\partial g}{\partial \lambda} + \frac{\partial g}{\partial \mu})/\sqrt{2}$ . We assume  $\mu \geq \lambda$  in the following proof,<sup>5</sup> since Donoho and Johnstone [20] have proven the part  $\mu < \lambda$ . Then simple calculations yield

$$\begin{aligned} h(\lambda, \mu) &= \sqrt{2} \cdot [(\Phi(\lambda + \mu) - \Phi(\mu - \lambda)) - \mu\varphi(\mu - \lambda) + \\ &\quad \varphi(\lambda + \mu)(\lambda^3 + 3\lambda^2\mu + 2\lambda\mu^2 + \mu - 2\lambda) - 1.2] \\ &\leq \sqrt{2}(0.5 + (\lambda + \mu)^3\varphi(\lambda + \mu) - 1.2) \\ &\leq (0.5 + 0.5 - 1.2) < 0. \end{aligned}$$

Therefore,

$$\rho_H(\lambda, \mu) \leq \min(\rho_H(\lambda, 0) + 1.2\mu^2, 1 + \lambda^2) \leq \min(2\varphi(\lambda)(\lambda + \frac{1}{\lambda}) + 1.2\mu^2, 1 + \lambda^2) \quad (2.50)$$

for any  $\lambda > 1$ . Now, combining (2.47) and (2.50) we can bound the univariate weighted lasso risk

$$\begin{aligned} \rho_W(\lambda, \mu) &\leq \max(\rho_S(\lambda, \mu), \rho_H(\lambda, \mu)) \leq (1 + \lambda_0^2) \min\left(\frac{2\varphi(\lambda_0)}{\lambda_0} + \frac{1.2}{1 + \lambda_0^2}\mu^2, 1\right) \\ &\leq (1 + \lambda_0^2) \min\left(\frac{2\varphi(\lambda_0)}{\lambda_0} + \mu^2, 1\right), \end{aligned}$$

for any  $\lambda_0 (= \lambda^{\frac{1}{1+\eta}}) > 1$ . And Theorem 3 follows.

Finally it may be worth mentioning that although applying Stein's lemma (see, for example, Gao [30]) is one possible way, it does not handle well the oracle bound for an estimator very close to hard thresholding — like Zou's oracle bound for the weighted lasso [61], after our correction mentioned in Section 2.2), because the hard-thresholding function is *not* weakly differentiable.

---

<sup>5</sup>In fact, with some more calculus we can prove that  $h(\lambda, \mu) \leq h(0, \sqrt{5}) \leq -0.008 < 0$  for any  $(\lambda, \mu) > 0$ .

## Chapter 3

# Sparse Regression with Exact Clustering

### 3.1 Background

In this chapter, we study a regression problem

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}), \quad (3.1)$$

where  $\mathbf{y}$  is the observed response vector,  $\mathbf{X}$  is the regression (design) matrix of size  $n$ -by- $p$ . Our task is to recover  $\boldsymbol{\beta}$  under some sparsity assumptions.

If  $\boldsymbol{\beta}$  is sparse in the (usual) sense that many of its components are zero (referred to as the zero-sparsity in this chapter), shrinkage methods are very attractive; among them, the lasso [50] is one of the most popular and basic ones. It is a convex optimization of

$$\min_{\boldsymbol{\beta}} \frac{1}{2} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1,$$

or stated in a multi-objective way [10]

$$\min(\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2, \|\boldsymbol{\beta}\|_1). \quad (3.2)$$

Keep in mind, however, that the test error is always our first concern in fitting a regression model – for example, the regularization parameter tuning is designed to serve this purpose, such as cross-validation. This is assumed throughout the chapter. One advantage of the

lasso is that it is computationally feasible to obtain an estimate with zero-sparsity.

Lasso has attracted a lot of attention recently because the nonsmooth  $l_1$ -penalty poses challenges in both fast computation for high-dimensional data and theoretical analysis of its performance, in terms of risk and sign consistency. These works aim to recover the zero-sparsity of  $\beta$  only. On the other hand, motivated by the intuition that the  $l_1$ -penalty is a convex approximation to the  $l_0$ -penalty in optimization, people have tried far more  $l_1$ -type penalties in fitting a regression model to capture various types of sparsity, especially in the field of signal processing. Unfortunately, there is not much theoretical analysis on their performances or a general algorithm efficient enough for solving the problem in high-dimensional space so far.

The chapter is organized as follows. Motivated by a gene clustering problem, Section 3.2 proposes a clustered lasso method, and more importantly, builds up a framework for the generic sparse regression with customizable sparsity patterns. A theoretical study is then performed in Section 3.3 on the power and limitations of the  $l_1$ -penalty in this framework, and improving techniques of data-augmentation and weights are investigated. Section 3.4 successfully tackles the computation problem of solving the sparse regression in high-dimensional space by developing an iterative algorithm with theoretical justifications. All technical details are left to the end of this chapter.

## 3.2 Clustered Lasso

The main purpose of this chapter is to do variable selection and variable grouping simultaneously. One of our motivations is based on a microarray study to discover the aging-related genes. This microarray dataset consists of large-scale gene expression data of 133 human kidney samples. The gene expression matrix  $\mathbf{X}$  is of size  $133 \times 44,928$ , and the responses,  $\mathbf{y}$ , are the ages of the 133 subjects.

Of course, we can run lasso to discover only a few genes as significant factors contributing to a regression model (3.1). But there is a serious inherent drawback of the lasso: it can select **at most**  $n$  predictors. Nonetheless, how to obtain more relevant variables in the model with just a few observations? A reasonable idea is to make the nonzero coefficients come out equal in clusters. Clearly, this keeps the model sparsity – it helps us achieve a *parsimonious* model in contrast to the ridge regression. It is more interpretative as a form of regularization, in terms of average expression values. It is also less sensitive to noise

because we can construct group-based predictors which have more accurate measurements. And later, these groups can be used to impute the missing values if data missing occurs. In consideration of these benefits, we would like to identify and group the relevant variables based on their effects, or coefficients.

The proposed problem requires combined regression and clustering analysis. One possible way is to directly apply some clustering approach to the estimated coefficients, which often results in an *ad-hoc* algorithm. The estimate in the first step may not be stable enough. And in doing the clustering in the second step, we need to specify a similarity measure and the number of clusters. Typically, the standard error information of the estimate is discarded in this step. More importantly, the clustering criterion is different than the test error, so the obtained clusters may not help in statistical modeling at all. For high-dimensional data, this two-step approach is unstable and inaccurate.

A more ambitious and more trustworthy way is to take the clusters into account when fitting the regression model, that is, we do the coefficient estimation and the coefficient clustering simultaneously. This can be achieved by integrating a penalty for improper clustering into the objective function. Indeed, one of the main tasks of this chapter is to develop a *sparse regression with exact clustering*. Why “exact”? In statistics, without the standard error information, one cannot determine how close two estimates, say,  $\hat{\beta}_i$  and  $\hat{\beta}_j$  are, even if the gap between them is small; however, if we get a gap estimate *exactly* equal to 0, its  $p$ -value would be 1, which means that we have enough confidence to put gene  $i$  and gene  $j$  into the same group.

In the language of multi-objective optimization [10], the problem can be formulated into

$$\min(\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2, \|\boldsymbol{\beta}\|_0, \sum_{i < j} 1_{\beta_i \neq \beta_j}). \quad (3.3)$$

Two types of sparsity are desirable: the *zero-sparsity* and the *equi-sparsity*, achieved by minimizing  $\|\boldsymbol{\beta}\|_0$  and  $\sum_{i < j} 1_{\beta_i \neq \beta_j}$ , respectively.

The problem of (3.3) is a combinatorial optimization and is NP-hard [1]. Motivated by the fact that the  $l_1$ -penalty is a convex approximation of the  $l_0$ -penalty in optimization, we may try to minimize

$$(\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2, \|\boldsymbol{\beta}\|_1, \sum_{i < j} |\beta_i - \beta_j|),$$

or equivalently,

$$\frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda_1\|\boldsymbol{\beta}\|_1 + \lambda_2\sum_{i<j}|\beta_i - \beta_j|, \quad (3.4)$$

after introducing two regularization parameters  $\lambda_1$  and  $\lambda_2$ . This will be referred to as the **clustered lasso**. Tuning  $\lambda_1$  and  $\lambda_2$  has a large influence on the algorithm performance. Usually, we do an empirical parameter search in the  $(\lambda_1, \lambda_2)$  space to minimize the validation error.

In what follows, we introduce a general framework for sparse regression. The goal is to obtain a regression estimate with  $\mathbf{T}$ -sparsity, i.e.,

$$\min(\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2, \|\mathbf{T}\boldsymbol{\beta}\|_0), \quad (3.5)$$

where  $\mathbf{T}$  is the sparsity pattern matrix specified by the user. A feasible alternative to overcome the NP-hardness is to solve the **sparse regression** of the form

$$\min(\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2, \|\mathbf{T}\boldsymbol{\beta}\|_1). \quad (3.6)$$

The  $\mathbf{T}$  matrix is used to characterize the *coding sparsity* in representing the true  $\boldsymbol{\beta}$ , not only the zero-sparsity in the usual sense. Here are some examples.

**Example 1. (Mixed  $\mathbf{T}$ )** If we know the successive differences of  $(\beta_1, \beta_2, \beta_3)$  are equal,  $\beta_3$  equals  $\beta_4$ , and  $\beta_5$  is zero, then  $\mathbf{T}$  may include rows

$$\begin{bmatrix} 1 & -2 & 1 & & \\ & & 1 & -1 & \\ & & & & 1 \end{bmatrix}$$

to capture this sparsity in fitting the regression model.

**Example 2. (Clustered/Fused lasso)** In our clustered lasso problem,

$$\mathbf{T} = \begin{bmatrix} \mathbf{I} \\ \lambda\mathbf{F} \end{bmatrix}, \quad (3.7)$$

where  $\mathbf{F}$  is a matrix including all pairwise differences (see (3.44)). And the fused lasso [51] replaces the  $\mathbf{F}$  in (3.7) by a neighboring difference matrix (see (3.42) for an illustration).

**Example 3. (Dense  $\mathbf{T}$ )**  $\mathbf{T}$  can be given by a wavelet transformation matrix (possibly overcomplete), which is useful in signal denoising and compression.

**Example 4. (Spatial  $\mathbf{T}$ )** In the field of computer vision and image processing, there exist many meaningful choices for  $\mathbf{T}$ , which can be constructed from various spatial operators, such as the following *Laplacian of Gaussian* used in edge detection

$$\begin{bmatrix} 0 & 1 & 0 \\ 1 & -4 & 1 \\ 0 & 1 & 0 \end{bmatrix}.$$

The two-dimensional fused lasso [27] also makes an example here.

In summary, the customizable  $\mathbf{T}$  represents the sparsity requirement posed by the user in the regression. Our studies in the rest of the chapter are made for an arbitrary  $\mathbf{T}$  matrix.

Though similar in form, the clustered lasso (3.4) is different than the fused lasso [51], in that it does not require the regression features to be ordered and so is much more challenging. In fact, it is trying to organize the unordered features and yields the ordering as an outcome, and, thus, can be used as a pre-processing step for the fused lasso. It might also be worthwhile to make a comparison between the clustered lasso and the grouped lasso methods [56, 59]. In grouped lasso, the grouping of features (predictors) is assumed to be known, arising naturally from the underlying background, such as the dummy variables introduced for a multi-level factor. Specifically, the coefficients from the same group need *not* be equal. By contrast, our clustered lasso is, indeed, a *supervised* clustering approach which looks for the appropriate grouping by taking both  $\mathbf{X}$  and  $\mathbf{y}$  into consideration.

Unfortunately, the (tuned) clustered lasso suffers some serious problems. First, its test error is often not small although it has two regularization parameters. More importantly, it barely shows enough proper clustering in our experiments. See Figure 3.1 for an illustration.

Somewhat surprisingly, this problem remains even after we tried a lot of ways to design the search scheme for choosing the values of the two regularization parameters. Increasing the sample size does not resolve the issue, either. Indeed, the widely acknowledged power of the  $l_1$  penalty to approximate the  $l_0$  penalty seems *specious* in this situation (as will be revealed by Theorem 4 below).

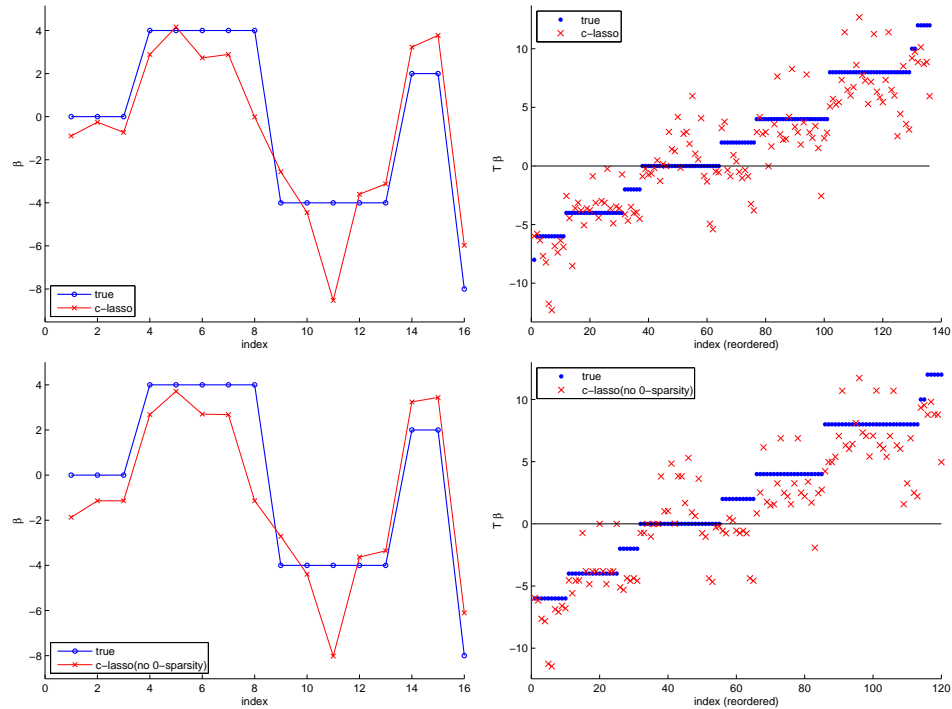


Figure 3.1: Clustered lasso does not show enough exact-clustering effect. The data is generated from the model described in Example 3 below. The upper left panel plots the (tuned) clustered lasso estimate  $\hat{\beta}$  and the true  $\beta$ , and the upper right panel compares  $T\hat{\beta}$  with  $T\beta$ , where the rows of  $T$  are reordered based on the values of  $T\beta$ . It is obvious that the clustered lasso can hardly capture all the true sparsity (corresponding to the part  $T\beta = \mathbf{0}$ , the indices given by  $38, \dots, 64$ ) in this example. The second row shows similar results if we drop the  $\|\beta\|_1$  in (3.4), which corresponds to a sparse regression problem seeking for equi-sparsity only.

### 3.3 Limitations and Improvements of the Clustered Lasso

#### 3.3.1 The power and limitations of the $L_1$ -penalty in sparse regression

It is widely known that the  $l_1$ -norm penalty is a convex approximation to the  $l_0$ -norm penalty in optimization. For instance, the variable selection in regression is a multi-objective problem with the  $l_0$ -penalty:

$$\min (\|\mathbf{y} - \mathbf{X}\beta\|_2^2, \|\beta\|_0).$$

Discovering the best subset of predictors is NP-hard [1]. Lasso, by replacing the  $l_0$ -norm with the  $l_1$ -norm in the criterion, offers a computationally feasible way to tackle this problem

in practice, although it is *not* always guaranteed to be sign consistent [60, 61]. For a general  $\mathbf{T}$ , the nature of this  $l_1$  approximation is worth more careful study in theory. As a result, we shall see why the naïve  $l_1$ -norm is not very successful for the clustered lasso although it is used in the same way as in the fused lasso.

As before [45], we adopt the generalized sign notation. Introduce  $\widetilde{\text{Sgn}}(\mathbf{v}) = \{\mathbf{s} : s_i = 1 \text{ if } v_i > 0, s_i = -1 \text{ if } v_i < 0, \text{ and } s_i \in [-1, 1] \text{ if } v_i = 0\}$ , and  $\widetilde{\text{sgn}}(\mathbf{v})$  is used to denote a specific element in  $\widetilde{\text{Sgn}}(\mathbf{v})$ . The usual sign vector is defined as  $\text{sgn}(\mathbf{v}) = \{\mathbf{s} : s_i = 1 \text{ if } v_i > 0, s_i = -1 \text{ if } v_i < 0, \text{ and } s_i = 0 \text{ if } v_i = 0\}$ .

Let  $\hat{\boldsymbol{\beta}}$  be an optimal solution to the generic sparse regression

$$\min_{\boldsymbol{\beta}} \frac{1}{2} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\mathbf{T}\boldsymbol{\beta}\|_1. \quad (3.8)$$

$\mathbf{T}$  may not have full rank. By the KKT optimality conditions [48] (the nonsmooth version),  $\hat{\boldsymbol{\beta}}$  is an optimal solution *if and only if*  $\hat{\boldsymbol{\beta}}$  satisfies

$$\mathbf{X}^T(\mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{y}) + \lambda \mathbf{T}^T \widetilde{\text{sgn}}(\mathbf{T}\hat{\boldsymbol{\beta}}) = \mathbf{0},$$

for some  $\widetilde{\text{sgn}}(\mathbf{T}\hat{\boldsymbol{\beta}})$ . We work in a classical setting ( $\mathcal{C}$ ): assume that  $p, \boldsymbol{\beta}$  are fixed and  $n \rightarrow \infty$ ;  $\boldsymbol{\Sigma} \triangleq \mathbf{X}^T \mathbf{X} / n \rightarrow \mathbf{C}$ , a positive definite matrix, with probability 1.

Throughout this chapter, given a matrix  $\mathbf{A}$ , we use  $\mathbf{A}_I$  to denote the submatrix of  $\mathbf{A}$  composed of the rows indexed by  $I$ , such that  $\mathbf{A}_I \boldsymbol{\alpha} = (\mathbf{A}\boldsymbol{\alpha})_I, \forall \boldsymbol{\alpha}$ . Given two matrices  $\mathbf{A}, \mathbf{B}$ ,  $\mathbf{B} \subset \mathbf{A}$  means that  $\mathbf{B}$  is a submatrix of  $\mathbf{A}$ , composed of some of its rows and all columns (so  $\mathbf{A}_I \subset \mathbf{A}$ ).

**Proposition 1** *If  $\lambda = o(n)$ , then  $\hat{\boldsymbol{\beta}} \xrightarrow{P} \boldsymbol{\beta}$ , and so  $\mathbf{T}\hat{\boldsymbol{\beta}} \xrightarrow{P} \mathbf{T}\boldsymbol{\beta}$ .*

Therefore consistency is a weak requirement, placing no restrictions on  $\boldsymbol{\Sigma}$  or  $\mathbf{T}$ . It can easily be achieved by a properly chosen  $\lambda$ . Yet in using the  $l_1$  penalty, we expect something more in sparsity recovery.

In this chapter, we also consider another notion of consistency useful in variable selection:

**Definition 1** (*Sign consistency [60]*) *Let  $\hat{\boldsymbol{\theta}}$  be a sequence of estimators of  $\boldsymbol{\theta}$ . Then  $\hat{\boldsymbol{\theta}}$  is defined to be sign-consistent if  $P(\text{sgn}(\hat{\boldsymbol{\theta}}) = \text{sgn}(\boldsymbol{\theta})) \rightarrow 1$ .*

Note that consistency implies nonzero sign consistency. For example, from Proposition 1,  $P(\text{sgn}(\hat{\boldsymbol{\beta}}_I) = \text{sgn}(\boldsymbol{\beta}_I)) \rightarrow 1$  for  $I = \{i : \boldsymbol{\beta}_i \neq 0\}$ , and  $P(\text{sgn}((\mathbf{T}\hat{\boldsymbol{\beta}})_I) = \text{sgn}((\mathbf{T}\boldsymbol{\beta})_I)) \rightarrow 1$

for  $I' = \{i : (\mathbf{T}\boldsymbol{\beta})_i \neq 0\}$ .

**Definition 2** (Zero  $s$ -consistency) Let  $\hat{\boldsymbol{\theta}}$  be a sequence of estimators of  $\boldsymbol{\theta}$  satisfying  $\mathbf{A}\boldsymbol{\theta} = \mathbf{0}$  for some matrix  $\mathbf{A}$ .  $\hat{\boldsymbol{\theta}}$  is defined to be zero  $s$ -consistent with respect to  $\mathbf{A}$  if  $P(\mathbf{A}\hat{\boldsymbol{\theta}} = \mathbf{0}) \rightarrow 1$ .

Note that here we care about the concentrated probability, but not the zero consistency in the usual sense. The zero- $s$  consistency is a key notion used to characterize sparsity recovery. For example, in the clustered lasso, zero- $s$  consistency means successfully discovering all the true groups asymptotically. Returning to our  $\mathbf{T}$ -sparsity problem, define  $z = z(\mathbf{T}, \boldsymbol{\beta}) = \{i : (\mathbf{T}\boldsymbol{\beta})_i = 0\}$ ,  $nz = nz(\mathbf{T}, \boldsymbol{\beta}) = \{i : (\mathbf{T}\boldsymbol{\beta})_i \neq 0\}$ .

**Proposition 2** If  $\lambda = O(\sqrt{n})$ , i.e.,  $\limsup_{n \rightarrow \infty} \lambda/\sqrt{n} < \infty$ , then  $\hat{\boldsymbol{\beta}}$  is not zero  $s$ -consistent with respect to  $\mathbf{T}_z$ .

In the following studies, a *joint* zero  $s$ -consistency will be our main concern. Namely, we study the conditions for zero  $s$ -consistency (w.r.t. some  $\mathbf{T}_1 \subset \mathbf{T}_z$ ) under the consistency assumption. (Note that the sign consistency for nonzero components follows from consistency.) This is because in practice, although blindly increasing  $\lambda$  would bring the thresholding power of the  $l_1$ -penalty into play, we prefer a value of  $\lambda$  with small test error, like one from cross-validation. The consistency requirement (weak!) corresponds to this way of tuning  $\lambda$ .

**Theorem 4** Assume the classical setup (C);  $\mathbf{T} = \begin{bmatrix} \mathbf{T}_1 \\ \mathbf{T}_2 \end{bmatrix}$ ,  $\mathbf{T}_1\boldsymbol{\beta} = \mathbf{0}$ ;  $\lambda/n \rightarrow 0$ ,  $\lambda/\sqrt{n} \rightarrow \infty$ .

We use  $\mathbf{A}^+$  to denote the Moore-Penrose inverse of  $\mathbf{A}$ . Then a necessary condition for  $\hat{\boldsymbol{\beta}}$  to be zero  $s$ -consistent w.r.t.  $\mathbf{T}_1$  is

$$\begin{aligned} \exists \widetilde{sgn}(\mathbf{T}_2\boldsymbol{\beta}) \text{ s.t. } & \|(\mathbf{T}_1\mathbf{C}^{-1}\mathbf{T}_1^T)^+(\mathbf{T}_1\mathbf{C}^{-1}\mathbf{T}_2^T) \cdot \widetilde{sgn}(\mathbf{T}_2\boldsymbol{\beta})\|_\infty \\ & \leq \|(\mathbf{T}_1\mathbf{C}^{-1}\mathbf{T}_1^T)^+(\mathbf{T}_1\mathbf{C}^{-1}\mathbf{T}_1^T)\|_\infty. \end{aligned} \quad (3.9)$$

And a sufficient condition is given by

$$\|(\mathbf{T}_1\mathbf{C}^{-1}\mathbf{T}_1^T)^+(\mathbf{T}_1\mathbf{C}^{-1}\mathbf{T}_2^T) \cdot \widetilde{sgn}(\mathbf{T}_2\boldsymbol{\beta})\|_\infty < 1, \forall \widetilde{sgn}(\mathbf{T}_2\boldsymbol{\beta}). \quad (3.10)$$

As a special case, suppose  $\mathbf{T}_z$  has full row rank; substitute  $\mathbf{T}_z$  for  $\mathbf{T}_1$ , and  $\mathbf{T}_{nz}$  for  $\mathbf{T}_2$ ,

then, (3.9) and (3.10) become

$$\|(\mathbf{T}_z \mathbf{C}^{-1} \mathbf{T}_z^T)^{-1} (\mathbf{T}_z \mathbf{C}^{-1} \mathbf{T}_{nz}^T) \cdot \text{sgn}(\mathbf{T}_{nz} \boldsymbol{\beta})\|_\infty \leq 1, \quad (3.11)$$

and

$$\|(\mathbf{T}_z \mathbf{C}^{-1} \mathbf{T}_z^T)^{-1} (\mathbf{T}_z \mathbf{C}^{-1} \mathbf{T}_{nz}^T) \cdot \text{sgn}(\mathbf{T}_{nz} \boldsymbol{\beta})\|_\infty < 1, \quad (3.12)$$

respectively. Therefore, the sufficient condition is pretty strong. Simple algebra shows that they further reduce to the irrepresentable conditions [60, 61] in the lasso case where  $\mathbf{T} = \mathbf{I}$ .

As another interesting example, suppose  $(\mathbf{T}_1, \mathbf{T}_2)$  is ‘separable’ in the sense that  $\mathbf{T} = \begin{bmatrix} \mathbf{T}_1 \\ \mathbf{T}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{T}_{11} & \mathbf{0} \\ \mathbf{0} & \mathbf{T}_{22} \end{bmatrix}$ . We write  $\mathbf{C}$  in a corresponding block form  $\begin{bmatrix} \mathbf{C}_1 & \mathbf{C}_{12} \\ \mathbf{C}_{12}^T & \mathbf{C}_2 \end{bmatrix}$ , and assume that  $\mathbf{C}_2$  is nonsingular. Then the LHS of (3.10) becomes

$$\|(\mathbf{T}_{11} \mathbf{S}^{-1} \mathbf{T}_{11}^T)^+ \mathbf{T}_{11} \mathbf{S}^{-1} \mathbf{C}_{12} \mathbf{C}_2^{-1} \mathbf{T}_{22}^T \cdot \widetilde{\text{sgn}}(\mathbf{T}_2 \boldsymbol{\beta})\|_\infty,$$

where  $\mathbf{S} = \mathbf{C}_1 - \mathbf{C}_{12} \mathbf{C}_2^{-1} \mathbf{C}_{12}^T$ . Therefore, if the entries of  $\mathbf{C}_{12}$  are small enough (in absolute value), the zero  $s$ -consistency w.r.t.  $\mathbf{T}_1$  naturally follows.<sup>1</sup> This is verified in the lasso study with  $\mathbf{T} = \mathbf{I}$ ; see [60, 45] for example. Unfortunately, clustered lasso does not fall into this category because the rows of the  $\mathbf{T}$  there encompass all pairwise differences and thus never result in a separable  $(\mathbf{T}_z, \mathbf{T}_{nz})$ .

This theorem indicates that in contrast to consistency, zero  $s$ -consistency imposes further constraints on  $\boldsymbol{\Sigma}$ , or the data, aside from the controllable regularization parameter  $\lambda$ . We will not go into the mathematical details here, but these conditions state that in general, we should have good control over  $(\mathbf{T}_z \mathbf{C}^{-1} \mathbf{T}_z^T)^+ (\mathbf{T}_z \mathbf{C}^{-1} \mathbf{T}_{nz}^T)$  to ensure  $l_1$ ’s power in discovering the true sparsity. In fact, if we consider the joint zero  $s$ -consistency w.r.t.  $\mathbf{T}_1$  on the signal class  $\Omega = \{\boldsymbol{\beta} : \mathbf{T}_1 \boldsymbol{\beta} = \mathbf{0}\}$ , the sufficiency condition (3.10) becomes

$$\|(\mathbf{T}_1 \mathbf{C}^{-1} \mathbf{T}_1^T)^+ (\mathbf{T}_1 \mathbf{C}^{-1} \mathbf{T}_2^T)\|_\infty < 1. \quad (3.13)$$

As a result, the magnitude of the entries of  $(\mathbf{T}_1 \mathbf{C}^{-1} \mathbf{T}_1^T)^+ \mathbf{T}_1 \mathbf{C}^{-1} \cdot \mathbf{T}_2^T$  plays an important role in capturing the  $\mathbf{T}_1$ -sparsity on  $\Omega$ . Given  $\boldsymbol{\beta}$ ,  $\mathbf{T}_1$ , and  $\mathbf{C}$ , this makes a big difference between the fused lasso and the clustered lasso: the  $\mathbf{T}_2$  of the latter contains up to  $O(p^2)$  more rows in addition to the  $\mathbf{T}_2$  of the first! Recall that the matrix infinity norm is the

<sup>1</sup>Note that  $(\mathbf{T}_{11} \mathbf{S}^{-1} \mathbf{T}_{11}^T)^+$  is a continuous function of  $\mathbf{C}_{12}$  since the rank of  $\mathbf{T}_{11} \mathbf{S}^{-1} \mathbf{T}_{11}^T$  is preserved.

maximum of the  $l_1$ -norms of all rows — the clustered lasso is certainly more likely to blow up the quantity on the LHS of (3.13)!

As an illustration, we use the previous example as shown in Figure 3.1 to compare the fused lasso (using the ordering from the true  $\beta$ ) and the clustered lasso. The data is generated from the model described in Example 3 below. For convenience, we ignore the zero-sparsity constraint in both. For (3.9) with  $\mathbf{T}_1$  replaced by  $\mathbf{T}_z$ , the LHS equals 0.6 and the RHS 1 in the fused lasso, but these quantities are 3.0 and 1.6, respectively, for the clustered lasso. In addition, the matrix infinity norm on the LHS of (3.13) is only 1.7 for the fused lasso, but 31.2 for the clustered lasso. Clearly, the fused lasso and the clustered lasso (although similar in form) show remarkable difference in the behavior of the  $l_1$ -penalty, the latter much more difficult to recover the true sparsity even asymptotically.

This explains the dilemma we encountered in the clustered lasso. No matter how we devise a plan to tune the regularization parameters, the design criterion is (and should be) reducing the generalization error approximated via, say, cross-validation. Therefore, the well-tuned regularization parameters cannot not be very large as seen from Proposition 1 (if we do not want our estimate to be inconsistent). But Proposition 2 and especially Theorem 4 limit the thresholding power of the  $l_1$ -penalty on the given data.

If it is fair to say that this requirement on  $\Sigma$  might not be very restrictive for the lasso, or even for the fused lasso, it becomes so stringent for the clustered lasso that the exact clustering effect expected from  $l_1$  is seldom seen strong enough in practice. As a result, we must think of ways to improve the naïve  $l_1$ -penalty to gain exact clustering.

### 3.3.2 Improving techniques

In the previous chapter (or see [45]), we studied how to improve lasso using data-augmentation and weights. We argued that lasso needs a data-augmentation modification, not only to resolve the ‘grouping’ issue [62] and the singularity problem, but also to help in effective variable selection. The elastic net [62] (eNet) was shown to be a special case. To further improve the sparsity, weights can be added in the  $l_1$  norm. In fact, it can be shown that asymptotically, weighted lasso can be sign consistent and enjoys the oracle properties (see Zou [61], as well as our further generalization [45]); in the nonasymptotic situation, it achieves a sharp oracle bound as soft- or hard-thresholding for orthogonal designs, one that significantly improves the (corrected) bound given in [61]. The simulation study showed

the strength of combining the data-augmentation and weights in lasso.

These techniques apply to the generic sparse regression (3.8). First, according to Theorem 4, although the naïve  $l_1$ -norm penalty cannot attain a sparse enough solution, ideally, if we could rescale the rows of  $\mathbf{T}$  in the following way

$$D\mathbf{T} = \begin{bmatrix} \mathbf{I} & \\ & \varepsilon\mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{T}_z \\ \mathbf{T}_{nz} \end{bmatrix} = \begin{bmatrix} \mathbf{T}_z \\ \varepsilon\mathbf{T}_{nz} \end{bmatrix} \triangleq \mathbf{T}',$$

then, the LHSs of (3.9) and (3.10) could be reduced significantly for  $\varepsilon$  small, while the RHSs remain unchanged. In fact, one of the advantages of the fused lasso is that the two tuned regularization parameters provide *two* adaptive weights for the absolute values of the components of  $\mathbf{T}\boldsymbol{\beta}$ . This weight construction, however, depends on the sparsity types; for a general  $\mathbf{T}$ , it may be infeasible to supply this extra information. Furthermore, it is really between the zero components (corresponding to  $\|\mathbf{T}_z\boldsymbol{\beta}\|_1$ ) and nonzero components (corresponding to  $\|\mathbf{T}_{nz}\boldsymbol{\beta}\|_1$ ) that the weights should make a big difference.

It is legitimate to think of adding more weights, or better assigning a weight to each individual absolute value term to handle the large nonzero components of  $\mathbf{T}\boldsymbol{\beta}$  in the penalization. In what follows, we give an asymptotic study on how to choose weights.

Given weights  $w_i$  (positive), consider the weighted sparse regression of the form

$$\min \frac{1}{2} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \sum w_i |(\mathbf{T}\boldsymbol{\beta})_i|. \quad (3.14)$$

**Theorem 5** *Assume the classical setup (C). Suppose*

$$w_i^{-1} = O_p(A(n)), \forall j \in z, \quad w_i = O_p(B(n)), \forall j \in nz.$$

*Then the optimal solution  $\hat{\boldsymbol{\beta}}$  to (3.14) is both zero  $s$ -consistent with respect to  $\mathbf{T}_z$  and  $\sqrt{n}$ -consistent (for some properly chosen  $\lambda(n)$ ), as long as*

$$A(n)B(n) \rightarrow 0. \quad (3.15)$$

*As an example,  $w_i^{-1} = O_p(1/\sqrt{n}), \forall j \in z$ , and  $w_i = O_p(1), \forall j \in nz$ .*

(3.15) is a broad condition: we only require the weights for the truly nonzero components

are at a smaller rate than the weights for the truly zero components, or essentially,

$$\max\{w_{nz}\}/\min\{w_z\} \xrightarrow{P} 0.$$

This provides a far more flexible way for weight construction. Although  $l_1$  is not capable of recovering the true sparsity directly, the theorem suggests that by use of a zero consistent weight sequence, it can achieve the zero  $s$ -consistency asymptotically. For example, we can choose  $1/w_i = |(\mathbf{T}\hat{\boldsymbol{\beta}}_{wts})_i|, \forall i$ , and, essentially, *any* consistent estimator  $\hat{\boldsymbol{\beta}}_{wts}$  is valid. (In fact, it does not even have to be an estimator of  $\boldsymbol{\beta}$ , which, in the lasso case, generalizes Theorem 2 in [61] and validates the one-step SCAD weights in [63], too.) On the other hand, one should be aware that Theorem 5 is only an asymptotic study with  $n \rightarrow \infty$  and  $p$  fixed, and the sufficiency condition given by (3.15) seems to be too loose in applications. In fact, in the nonasymptotic situation, the weight choices, as observed in most applications, do have a large impact on the resulting sparsity and test error, which cannot be deduced from the theorem. And the notion of (large  $n$ ) consistency loses much of its meaning in our microarray data analysis, where  $p \gg n$ . In all, the weighted  $l_1$ -penalty is one way to increase model sparsity, but careful consideration needs to be given to the practical weight construction.

On the other hand, adding weights typically does not help improve the test error in practice; sometimes it can hurt the goodness-of-fit to some extent. This is undesirable in statistical modeling and explains why we want to combine weights with the data augmentation technique. In fact, even the data-augmentation with a not-so-good estimator may effectively reduce the test error, as seen in the eNet which uses  $\hat{\boldsymbol{\beta}}_{indv} = [\mathbf{x}_i^T \mathbf{y} / \mathbf{x}_i^T \mathbf{x}_i]_{p \times 1}$  [45] (not even consistent in the nonorthogonal case). In addition, we need data-augmentation to deal with the high singularity that inevitably arises in fitting a model to the microarray data with  $p \gg n$ . The  $l_1$  constraint, though capable of yielding a solution for  $p > n$ , thus, alleviating the singularity problem, still suffers from the nonuniqueness in some cases [62]; in contrast, data-augmentation decorrelates the columns of the design matrix by augmenting  $\mathbf{X}$  (diagonally or nondiagonally – see the last chapter for more details) and, therefore, perfectly resolves the issue. Recall that the data-augmentation technique (see Chapter 2 or [45]) refers to replacing  $\mathbf{X}^T$  and  $\mathbf{y}^T$  by  $\begin{bmatrix} \mathbf{X}^T & \mathbf{X}_{aug}^T \end{bmatrix}$  and  $\begin{bmatrix} \mathbf{y}^T & \mathbf{y}_{aug}^T \end{bmatrix}$ , respectively, say, by use of a known estimate  $\hat{\boldsymbol{\beta}}_{aug}$ . We can do it diagonally, taking  $\mathbf{X}_{aug} = \tau \mathbf{I}$  and

$\mathbf{y}_{aug} = \tau \hat{\boldsymbol{\beta}}_{aug}$ , as seen in eNet, or nondiagonally,  $\mathbf{X}_{aug} = \tau(\mathbf{I} - \hat{\boldsymbol{\beta}}_{aug} \hat{\boldsymbol{\beta}}_{aug}^T / \|\hat{\boldsymbol{\beta}}_{aug}\|_2^2)$  and  $\mathbf{y}_{aug} = \mathbf{0}$ . The latter is proposed in the last chapter, as an outcome of introducing a data-dependent scale parameter to improve the eNet. It is more robust to a bad  $\hat{\boldsymbol{\beta}}_{aug}$ . In the next subsection, we will give some specific designs of how to make use of these techniques together to reduce test error and increase model sparsity at the same time.

### 3.3.3 Algorithm design and a simulation study

In this part, we discuss the practical algorithm design and illustrate the strength of the data-augmented weighted (DAW) sparse regression by a simulation study.

First, (3.8) can be reparameterized assuming that  $\mathbf{T}$  has full row rank: introducing  $\mathbf{H}$  with  $\mathbf{HT} = \mathbf{I}$ ,  $\mathbf{Z} = \mathbf{XH}$ , the generic sparsity regression problem (3.8) is then equivalent to

$$\min_{\boldsymbol{\gamma}} \frac{1}{2} \|\mathbf{y} - \mathbf{Z}\boldsymbol{\gamma}\|_2^2 + \lambda \|\boldsymbol{\gamma}\|_1 \quad \text{s.t. } \mathbf{TH}\boldsymbol{\gamma} = \boldsymbol{\gamma}. \quad (3.16)$$

And the optimal  $\boldsymbol{\beta}_{opt}$  is obtained from the optimal  $\boldsymbol{\gamma}_{opt}$  by  $\boldsymbol{\beta}_{opt} = \mathbf{H}\boldsymbol{\gamma}_{opt}$ . This is a constrained lasso problem in the  $\mathbf{Z}$ -domain.

Given  $\hat{\boldsymbol{\beta}}_{aug}$ , we perform a *nondiagonal* data-augmentation [45] in the  $\mathbf{Z}$ -domain. Together with the weights constructed via some  $\hat{\boldsymbol{\beta}}_{wts}$ , this amounts to solving the following optimization problem in the  $\mathbf{X}$ -domain:

$$\min_{\boldsymbol{\beta}} \frac{1}{2} \left\| \begin{bmatrix} \mathbf{y} \\ \mathbf{y}_{aug} \end{bmatrix} - \begin{bmatrix} \mathbf{X} \\ \mathbf{X}_{aug} \end{bmatrix} \boldsymbol{\beta} \right\|_2^2 + \lambda \sum w_i |(T\boldsymbol{\beta})_i|, \quad (3.17)$$

where  $\mathbf{X}_{aug} = \sqrt{\tau} \mathbf{T}(\mathbf{I} - \hat{\boldsymbol{\beta}}_{aug} \hat{\boldsymbol{\beta}}_{aug}^T \mathbf{T}^T \mathbf{T} / \|\mathbf{T} \hat{\boldsymbol{\beta}}_{aug}\|_2^2)$ ,  $\mathbf{y}_{aug} = \mathbf{0}$ ,  $w_i = 1/|(\mathbf{T} \hat{\boldsymbol{\beta}}_{wts})_i|$ , and  $\lambda, \tau$  are two regularization parameters to be tuned. The general form of (3.17) will be referred to as the DAW version of the sparse regression; particularly, the DAW version of the clustered lasso with  $\mathbf{T}$  in the form of (3.7) will be called DAW-CLASSO for convenience. Note that unlike the clustered lasso or the fused lasso, the DAW version of  $\mathbf{T}$  has no need to include a regularization parameter inside.

The parameter tuning plays an important role in determining an algorithm's performance, especially for those with multiple regularization parameters. In our simulations, we use the *alternative* search strategy (see Section 2.3) which has been shown to be fast and efficacious. Note that this is performed for two equivalent regularization parameters at each

step, but not always  $\lambda$  and  $\tau$ .

We develop two methods to solve our exact-clustering problem.

### Method A

In method A, two initial estimates are used – the ranged lasso estimate  $\hat{\beta}_{r-lasso}$ , and the clustered lasso estimate  $\hat{\beta}_{c-lasso}$  (from (3.4)). The ranged lasso was proposed in the last chapter and has been shown to be a competitive alternative to ridge regression with consistently smaller test error (and some sparsity, too). It provides a good choice for  $\hat{\beta}_{aug}$ . In the setting of the sparse regression (3.16), we first get the (constrained) ranged lasso solution  $\hat{\gamma}_{r-lasso}$  in the  $\mathbf{Z}$ -domain and then project it back to get  $\hat{\beta}_{r-lasso}$ . Method A goes as follows.

Step 1. Fit a DAW-CLASSO model by substituting  $\hat{\beta}_{r-lasso}$  for  $\hat{\beta}_{aug}$ , and  $\hat{\beta}_{c-lasso}$  for  $\hat{\beta}_{wts}$  in (3.17).

Step 2. Fit the DAW-CLASSO again, but use the last estimate as both  $\hat{\beta}_{aug}$  and  $\hat{\beta}_{wts}$ .

### Method B

Method B, in contrast, does not start from a good initial estimate. It fits a fused lasso model as long as the ordering is available, given by the clustered lasso or its DAW version.

Step 1. Use the ordering got from  $\hat{\beta}_{c-lasso}$  (cf. (3.4)) to fit a fused lasso model, the estimate denoted by  $\hat{\beta}_{f-lasso}$ .

Step 2. Fit a DAW-CLASSO by substituting  $\hat{\beta}_{c-lasso}$  for  $\hat{\beta}_{aug}$ , and  $\hat{\beta}_{f-lasso}$  for  $\hat{\beta}_{wts}$ , the estimate denoted by  $\hat{\beta}_{daw-lasso}$ .

Step 3. Repeat Step 1 using the new ordering obtained from the last estimate, getting a new fused lasso estimate.

Step 4. Repeat Step 2 with  $\hat{\beta}_{daw-lasso}$  for data-augmentation and the updated fused lasso estimate for weight construction.

We did experiments on three simulation datasets. Each dataset contains training data, validation data, and test data. We use  $\# = \cdot / \cdot / \cdot$  to denote the number of observations in the training data, validation data, and test data. Let  $\Sigma$  be the correlation matrix in generating  $\mathbf{X}$ , i.e., each row of  $\mathbf{X}$  is independently drawn from  $N(\mathbf{0}, \Sigma)$ . We use

$(\{a_1\}^{n_1}, \dots, \{a_k\}^{n_k})$  to denote the column vector made by  $n_1$   $a_1$ 's,  $\dots$ ,  $n_k$   $a_k$ 's consecutively in the following examples.

**Example 1.**  $\# = 20/100/100$ ,  $d = 13$ ,  $\beta = (\{0\}^2, \{-1.5\}^2, \{-2\}^2, \{0\}^2, \{1\}^2, \{4\}^3)$ ,  $\sigma = 5$ ,  $\Sigma_{ij} = \rho^{|i-j|}$  with  $\rho = 0.5$ .

**Example 2.**  $\# = 20/100/100$ ,  $d = 13$ ,  $\beta = (\{0\}^2, \{-1.5\}^2, \{-2\}^2, \{0\}^2, \{1\}^2, \{4\}^3)$ ,  $\sigma = 5$ ,  $\Sigma_{ij} = \rho^{|i-j|}$  with  $\rho = 0.9$ .

**Example 3.**  $\# = 30/100/100$ ,  $d = 16$ ,  $\beta = (\{0\}^3, \{4\}^5, \{-4\}^5, \{2\}^2, \{-8\}^1)$ ,  $\sigma = 5$ ,  $\Sigma_{ij} = (-1)^{(i-j)} \cdot 0.8$  for  $i \neq j$ .

Example 1 and Example 2, the second much more correlated than the first, demonstrate a situation of many small clusters in  $\beta$ , where overlap is likely to occur. In the third example, big clusters coexist with a few small clusters and the signal to noise variance ratio is low. Before an algorithm is applied, the columns of a regression matrix are all normalized to have a squared  $l_2$ -norm equal to the number of the observations; no centering is performed in these examples.

Each model is simulated 50 times; then, we measure the performance of each algorithm by the test error and the proper sparsity. The test error is characterized by the 40% trimmed-mean<sup>2</sup> of the scaled MSE (SMSE) on the test data, where SMSE is  $100 \cdot (\sum_{i=1}^N (\hat{y}_i - y_i)^2 / (N\sigma^2) - 1)$  defined for the test data. The proper sparsity here is defined by the 40% trimmed-mean of the following 50 percentages:  $100\% \cdot |\{i : (\mathbf{T}_z \hat{\beta})_i = 0\}| / |z|$ , which represents the number of proper zeros for each estimate.

Seen from the table, the clustered lasso does not exhibit enough exact-clustering in these examples, and its test errors are not satisfactory. By contrast, the ranged lasso provides a more accurate estimate although it may not be very sparse. Owing to the DAW, both Method A and Method B significantly improve the performance of the clustered lasso: the test error is reduced effectively, as a result of (nondiagonal) data-augmentation; *simultaneously*, the proper sparsity is enhanced by use of the weights, by about 50% at the minimum. The results are very encouraging. DAW-improved  $l_1$  helps us get a good model; at the same time, it accomplishes the variable selection and the variable grouping.

<sup>2</sup>Medians of errors are mostly used [50, 62] to measure the algorithm performance from multiple runs, but are not so stable for comparisons based on our experience. Discarding 20 highest and 20 lowest errors, we compute the average of the remaining 10.

	Example 1		Example 2		Example 3	
	Test-err	p-Spar	Test-err	p-Spar	Test-err	p-Spar
c-lasso	<b>45.0</b>	<b>15.8%</b>	<b>22.1</b>	<b>22.1%</b>	<b>69.3</b>	<b>5.3%</b>
r-lasso	39.3	8.6%	18.3	12.2%	64.6	1.5%
A-step 1	34.8	22.5%	14.3	31.6%	61.0	12.8%
A-step 2	<b>35.8</b>	<b>23.9%</b>	<b>16.7</b>	<b>36.7%</b>	<b>62.2</b>	<b>15.1%</b>
B-step 1	45.3	26.0%	20.9	38.2%	69.5	16.4%
B-step 2	40.2	30.9%	16.3	39.2%	63.2	10.5%
B-step 3	37.3	30.7%	18.9	38.0%	68.4	18.2%
B-step 4	<b>35.3</b>	<b>30.8%</b>	<b>15.0</b>	<b>37.5%</b>	<b>60.5</b>	<b>12.5%</b>

Table 3.1: Performance comparisons on the simulation data, in terms of test error – 40% trimmed-mean SMSE, and proper sparsity – 40% trimmed-mean of percentages of proper zeros in the estimates. There, c-lasso refers to the clustered lasso estimate, and r-lasso the ranged lasso estimate; Method A uses r-lasso to run DAW-CLASSO, while Method B fits a fused lasso model in step 1 and step 3, with the ordering extracted from the clustered lasso or its DAW version.

### 3.4 A Fast Algorithm for Solving the Generic Sparsity Problem

In applying the (improved) clustered lasso to the microarray data, we encounter insurmountable difficulty with all the optimization procedures (to date), largely due to the fact that  $\mathbf{T}$  has  $O(d^2)$  rows: our experience shows that it is extremely difficult or infeasible to carry out the supervised clustering for  $d > 100$ . In this section, we propose a simple but fast algorithm to solve the generic sparsity problem in applications with large data.

#### 3.4.1 Motivation

Let's first review the canonical lasso problem: given  $(\mathbf{X}, \mathbf{y}, \lambda)$ ,

$$\min \frac{1}{2} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1. \quad (3.18)$$

By the KKT optimality conditions [48] (the nonsmooth version),  $\hat{\boldsymbol{\beta}}$  is an optimal solution *if and only if*  $\hat{\boldsymbol{\beta}}$  satisfies the equation

$$\mathbf{X}^T(\mathbf{X}\boldsymbol{\beta} - \mathbf{y}) + \lambda \widetilde{\text{sgn}}(\boldsymbol{\beta}) = \mathbf{0}, \text{ or } \lambda \widetilde{\text{sgn}}(\boldsymbol{\beta}) = \mathbf{X}^T \mathbf{y} - \boldsymbol{\Sigma} \boldsymbol{\beta}, \quad (3.19)$$

where the same generalized sign notation (not a function) —  $\widetilde{\text{sgn}}$  [45], is used, to denote a subgradient of  $\|\boldsymbol{\beta}\|_1$ . It is easy to see the following important fact about  $\widetilde{\text{sgn}}$ :

Fact) Given an arbitrary  $\widetilde{\text{sgn}}(\boldsymbol{\beta}) \in \widehat{\text{Sgn}}(\boldsymbol{\beta})$ , let  $\boldsymbol{\xi} = \boldsymbol{\beta} + \lambda \widetilde{\text{sgn}}(\boldsymbol{\beta})$ , then

$$\boldsymbol{\beta} = \Theta_S(\boldsymbol{\xi}; \lambda),$$

where  $\Theta_S(\cdot; \lambda)$  (or  $\Theta(\cdot; \lambda)$ , for simplicity) is the soft-thresholding operator using  $\lambda$  as the threshold value.

Rewrite (3.19) as

$$\boldsymbol{\beta} + \lambda \widetilde{\text{sgn}}(\boldsymbol{\beta}) = \mathbf{X}^T \mathbf{y} + (\mathbf{I} - \boldsymbol{\Sigma})\boldsymbol{\beta}. \quad (3.20)$$

This motivates an iterative design to solve (3.18)

$$\boldsymbol{\xi}^{(j+1)} = \mathbf{X}^T \mathbf{y} + (\mathbf{I} - \boldsymbol{\Sigma})\boldsymbol{\beta}^{(j)}, \quad \boldsymbol{\beta}^{(j+1)} = \Theta(\boldsymbol{\xi}^{(j+1)}; \lambda). \quad (3.21)$$

If  $\|\boldsymbol{\Sigma}\|_2 < 1$ , this nonlinear process can be shown to *converge*<sup>3</sup> to an optimal point even if  $\boldsymbol{\Sigma}$  is singular (in which case  $(\mathbf{I} - \boldsymbol{\Sigma})$  is not a contraction, but only *nonexpansive*) [18]. The corresponding algorithm is simple and does not involve matrix inverse computation.

The iteration of (3.21) has been proposed in different forms and is advocated by many researchers [18, 27, 55] in large-data problems (even competitive with LARS). In contrast to our starting point of the generalized sign ( $\widetilde{\text{sgn}}$ ), Daubechies *et al.* [18] constructed it from surrogate functions, Friedman *et al.* [27], Wu and Lange [55] are based on coordinate optimization. In particular, Daubechies *et al.* [18] prove nice theoretical results on its convergence in a general functional framework; Friedman *et al.* [27] demonstrate the amazing performance of this iterative algorithm in terms of the computation time compared to the homotopy method and LARS.

Now consider the generic sparsity problem

$$\min \frac{1}{2} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\mathbf{T}\boldsymbol{\beta}\|_1, \quad (3.22)$$

where  $\mathbf{T}$  is a sparsity pattern matrix to be specified by users in different situations. Throughout this section, we assume  $\mathbf{T}$  has full column rank,<sup>4</sup> which means that it is a square or

<sup>3</sup>Note that this theoretical achievement is considerably **stronger** than an ‘every accumulation point’ argument often seen (e.g., in [7]).

<sup>4</sup>Otherwise, it is no more difficult than the nonsingular (square) case.

‘thin’ matrix. The optimal  $\hat{\beta}$  satisfies the equation

$$\mathbf{X}^T(\mathbf{X}\beta - \mathbf{y}) + \lambda \mathbf{T}^T \widetilde{\text{sgn}}(\mathbf{T}\beta) = \mathbf{0},$$

or

$$\mathbf{T}^T \cdot \Theta(\mathbf{T}\beta; \lambda) = \mathbf{X}^T \mathbf{y} + (\mathbf{T}^T \mathbf{T} - \Sigma)\beta.$$

The difficulty is, however,  $\mathbf{T}^T$  has no left inverse in the case of a ‘thin’  $\mathbf{T}$ . For example, in the fused lasso,

$$\mathbf{T} = \begin{bmatrix} \mathbf{I} \\ \lambda_2 \mathbf{F} \end{bmatrix} \quad \text{with } \mathbf{F} = \begin{bmatrix} 1 & -1 & & & \\ & 1 & -1 & & \\ & & \dots & \dots & \\ & & & & 1 & -1 \end{bmatrix}.$$

$\mathbf{T}$  does not have a right inverse. Consequently, we are incapable of developing a proper iterative equation to get the exact solution.

The above argument explains why Friedman, Hastie, and Tibshirani [27] encountered difficulty in generalizing the coordinate optimization from the lasso to the fused lasso. Introducing a ‘descent cycle’, a ‘fusion cycle’, and a ‘smoothing cycle’, Friedman *et al.* designed an *ad-hoc* algorithm for solving the diagonal fused lasso. There is seemingly no theoretical guarantee for its convergence or converging to an optimal solution.

We reparameterize (3.22). First, introduce  $\mathbf{H}$  with  $\mathbf{HT} = \mathbf{I}$ : assuming that the SVD decomposition of  $\mathbf{T}$  is given by  $\mathbf{T} = \mathbf{UDV}^T$ , we take  $\mathbf{H} = \mathbf{VD}^{-1}\mathbf{U}^T$  throughout this section. The generic sparsity regression problem (3.22) is equivalent to the following *constrained* lasso problem:

$$\min f(\gamma) \triangleq \frac{1}{2} \|\mathbf{y} - \mathbf{XH} \cdot \gamma\|_2^2 + \lambda \|\gamma\|_1 \quad \text{s.t. } \mathbf{TH}\gamma = \gamma. \quad (3.23)$$

The optimal  $\beta_{opt}$  is obtained from the optimal  $\gamma_{opt}$  by  $\beta_{opt} = \mathbf{H}\gamma_{opt}$ . It suggests an iterative algorithm as follows

$$\begin{cases} \gamma^{(j)} = \Theta(\mathbf{H}^T \mathbf{X}^T \mathbf{y} + (\mathbf{I} - \mathbf{H}^T \Sigma \mathbf{H})\gamma^{(j-1)}; \lambda) \\ \gamma^{(j+1)} = \mathbf{TH}\gamma^{(j)} \end{cases} \quad (3.24)$$

and

$$\boldsymbol{\beta}^{(j)} = \mathbf{H}\boldsymbol{\gamma}^{(j)}. \quad (3.25)$$

The good news is that this iteration, though only *nonexpansive* for  $\boldsymbol{\Sigma}$  singular, does converge<sup>5</sup> (under some mild conditions); the bad news is that it does not converge to the right point. Nevertheless, (3.24) is the basic iteration of our practical algorithms below.

### 3.4.2 The ‘annealing’ algorithm

Observe that the original optimization problem (3.22) is also equivalent to

$$\min \frac{1}{2} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \frac{\lambda}{k} \|(\mathbf{T} \cdot k)\boldsymbol{\beta}\|_1.$$

for *any*  $k$  positive. We may use a variant of (3.24) and (3.25)

$$\begin{cases} \tilde{\boldsymbol{\gamma}}^{(j)} = \Theta \left( \frac{1}{k} \mathbf{H}^T \mathbf{X}^T \mathbf{y} + \left( \mathbf{I} - \frac{1}{k^2} \mathbf{H}^T \boldsymbol{\Sigma} \mathbf{H} \right) \tilde{\boldsymbol{\gamma}}^{(j-1)}; \frac{\lambda}{k} \right) \\ \tilde{\boldsymbol{\gamma}}^{(j+1)} = \mathbf{T} \mathbf{H} \tilde{\boldsymbol{\gamma}}^{(j)}, \end{cases} \quad (3.26)$$

$$\boldsymbol{\beta}^{(j)} = \mathbf{H} \tilde{\boldsymbol{\gamma}}^{(j)} / k. \quad (3.27)$$

Let  $\boldsymbol{\gamma}^{(j)} = \tilde{\boldsymbol{\gamma}}^{(j)} / k$ . Since

$$\tilde{\boldsymbol{\gamma}}^{(j)} + \frac{\lambda}{k} \widetilde{\text{sgn}}(\tilde{\boldsymbol{\gamma}}^{(j)}) = k\boldsymbol{\gamma}^{(j)} + \frac{\lambda}{k} \widetilde{\text{sgn}}(\boldsymbol{\gamma}^{(j)} \cdot k) = k\boldsymbol{\gamma}^{(j)} + \frac{\lambda}{k} \widetilde{\text{sgn}}(\boldsymbol{\gamma}^{(j)}),$$

we get

$$\begin{cases} \boldsymbol{\gamma}^{(j)} = \Theta \left( \frac{1}{k^2} \mathbf{H}^T \mathbf{X}^T \mathbf{y} + \left( \mathbf{I} - \frac{1}{k^2} \mathbf{H}^T \boldsymbol{\Sigma} \mathbf{H} \right) \boldsymbol{\gamma}^{(j-1)}; \frac{\lambda}{k^2} \right) \\ \boldsymbol{\gamma}^{(j+1)} = \mathbf{T} \mathbf{H} \boldsymbol{\gamma}^{(j)}, \end{cases} \quad (3.28)$$

$$\boldsymbol{\beta}^{(j)} = \mathbf{H} \boldsymbol{\gamma}^{(j)}. \quad (3.29)$$

---

<sup>5</sup>Yet Daubechies *et al.*'s convergence theorem [18] cannot be directly applied, because  $\|\mathbf{H}^T \boldsymbol{\Sigma} \mathbf{H}\|_2$  is exactly one, but not less than one.

For clarity, we write (3.28) as even and odd updates

$$\begin{cases} \gamma_e^{(j)} = \Theta\left(\frac{1}{k^2}\mathbf{H}^T\mathbf{X}^T\mathbf{y} + \left(\mathbf{I} - \frac{1}{k^2}\mathbf{H}^T\boldsymbol{\Sigma}\mathbf{H}\right)\gamma_o^{(j-1)}; \frac{\lambda}{k^2}\right), \\ \gamma_o^{(j)} = \mathbf{T}\mathbf{H}\gamma_e^{(j)}. \end{cases} \quad (3.30)$$

**Theorem 6** *The following results hold for the iteration (3.30):*

1. *Convergence.* There exists a  $k_0 > 0$  such that for any  $k > k_0$ ,  $\gamma_e^{(j)}$ ,  $\gamma_o^{(j)}$ ,  $\beta^{(j)}$  converge given any initial value in (3.30). That is, as  $j \rightarrow \infty$ , we have

$$\gamma_e^{(j)}(k) \rightarrow \gamma_e(k), \gamma_o^{(j)}(k) \rightarrow \gamma_o(k), \beta^{(j)}(k) \rightarrow \beta(k).$$

2. *Optimality.* As  $k \rightarrow \infty$ , every limit point of  $\beta(k)$  (or  $\gamma_e(k)$ ,  $\gamma_o(k)$ ) is an optimal solution to (3.22) (or (3.23)).
3. *Rate.* Let  $\Delta(k) = \gamma_e(k) - \gamma_o(k)$ ,  $f_{\text{opt}}$  be the optimal value in (3.23). Then,<sup>6</sup>

$$\|\Delta(k)\| \leq \frac{C}{k^2}, \quad (3.31)$$

and

$$0 \leq f_{\text{opt}} - f(\gamma_e(k)) \leq \frac{C}{k^2}. \quad (3.32)$$

4.  *$k_0$ .* Finally,

$$k_0 \leq \frac{1}{\sqrt{2}} \frac{\sigma_{\max}(\mathbf{X})}{\sigma_{\min}(\mathbf{T})}, \quad (3.33)$$

where  $\sigma_{\max}(\sigma_{\min})$  denotes the largest (smallest) singular value of the corresponding matrix.

See Section 3.7 for the details of the proof, where we successfully generalize Daubechies *et al.*'s convergence theorem [18]. In the following, we abbreviate the subscripts of  $\gamma_e^{(j)}(k)$  and  $\gamma_e(k)$  for simplicity. We summarize more findings in the case that  $\boldsymbol{\Sigma}$  is nonsingular:

---

<sup>6</sup>In this chapter, we use  $C$  to denote a positive constant; yet these  $C$ 's may not take the same value, even in a single equation.

**Proposition 3** *Suppose  $\Sigma$  is nonsingular. Then*

$$\gamma_{opt} \triangleq \arg \left( \min_{\gamma} f(\gamma) \text{ s.t. } \mathbf{T}\mathbf{H}\gamma = \gamma \right)$$

*is unique. On the convergence of  $\gamma^{(j)}(k)$  ( $k > k_0$ ), we have*

$$\|\gamma^{(j)}(k) - \gamma(k)\| \leq \left(1 - \frac{\rho_0}{k^2}\right)^j \|\gamma^{(0)}(k) - \gamma(k)\|, \quad (3.34)$$

*where  $\rho_0 = \lambda_{\min}^+(\mathbf{H}^T \Sigma \mathbf{H})$ , the smallest positive eigenvalue of  $\mathbf{H}^T \Sigma \mathbf{H}$ ; and*

$$\|\gamma(k) - \gamma_{opt}\| \leq \frac{C}{k^2}. \quad (3.35)$$

*Moreover, sign consistency is achieved for finite  $k$ , — in particular,*

$$(\gamma(k))_z = \mathbf{0},$$

*for any finite  $k$  large enough, with the index set  $z$  satisfying  $(\gamma_{opt})_z = \mathbf{0}$ .*

From (3.34), with  $\delta \triangleq \|\gamma^{(0)}(k) - \gamma(k)\|$ , we have  $\|\gamma^{(j)}(k) - \gamma(k)\| \leq \epsilon_0$  if  $(1 - \frac{\rho_0}{k^2})^j \delta \leq \epsilon_0$  or  $j \leq \frac{\log(\delta/\epsilon_0)}{\log(1-\rho_0/k^2)} \approx k^2 \cdot \frac{1}{\rho_0} \log(\delta/\epsilon_0)$ , which indicates that the number of iterations required at  $k$  is  $O(k^2)$ ; on the other hand, from (3.31), or (3.35), the error  $\propto \frac{1}{k^2}$ . This is true in general: for a small  $k$ ,  $\beta^{(j)}(k)$  converges faster, but to an inaccurate solution; for a large  $k$ ,  $\beta^{(j)}(k)$  converges more slowly, but to a more accurate point.

Therefore we design an ‘annealing’ algorithm (not simulated annealing) with  $k$  acting as the inverse temperature parameter. Run (3.30) for small  $k$  first, then feed the estimate as the initial value into a new iteration associated with a larger  $k$ . The outline for our annealing algorithm is given as follows. The details of the design are given in the next subsection.

1. Initialization. Set the starting values for  $\gamma_o^{(0)}$ ,  $k$ , etc.
2. Iteration.
  - Update  $\gamma_e^{(j)}$ ,  $\gamma_o^{(j)}$ , and  $\beta^{(j)}$  in the way of (3.30) and (3.29).
  - If  $\|\beta^{(j)} - \beta^{(j-1)}\|$  is ‘small’, then
    - If  $\|\gamma_e^{(j)} - \gamma_o^{(j)}\|$  is ‘small enough’, exit.

- Otherwise, increase  $k$  to some ‘larger’ value.
- Let  $j \leftarrow j + 1$ ; go to the next iteration.

Both the inner  $j$ -convergence and the outer  $k$ -convergence can be slow: typically, they may fail to be linear (or geometric) – see (3.35) for an example; this is caused by the non-expansive operators. We need to think of some effective ways to boost both convergences to solve the problems in the real world.

### 3.4.3 Accelerated annealing

It is natural to think of updating  $k$  at each iteration  $j$ . Like simulated annealing (SA), in this *inhomogeneous* way, the ‘cooling schedule’, i.e., the growing manner of  $k(j)$ , is crucial to guarantee an optimal convergent point which solves (3.23).

**Theorem 7** *Assume  $\Sigma$  is nonsingular. If  $k(j)$  satisfies*

$$\sum_{j=1}^{\infty} \frac{1}{k^2(j)} = \infty, \text{ and } k(j) \rightarrow \infty \text{ as } j \rightarrow \infty, \quad (3.36)$$

*then the inhomogeneous chain must converge to the unique optimal solution.*

For example, we can take  $k(j) = \sqrt{j}$ . A detailed proof of Theorem 7 is provided in Section 3.7, based on a useful decomposition for inhomogeneous chains, due to Wrinkler [54]. In general, a valid cooling schedule, in theory, should be no faster than the  $k(j)$  satisfying (3.36).

Theorem 7 implies that it essentially takes polynomial time to yield a good solution, in contrast to the exponential in SA [40]. But  $\sqrt{j}$  might still be too slow in practice. In most applications, we are only interested in obtaining a good enough solution, which allows for an even faster cooling schedule. Based on our experience, we would recommend the *homogenous* updating – run a sequence of homogenous chains, each at a fixed  $k$ ; the trick here is to run the chains for small  $k$ ’s first, to complete the major improvements over the initial point, but not for too long a time, while the finer tunings are left to larger  $k$ ’s to aim better at  $\gamma_{opt}$ , rather than the inaccurate  $\gamma(k)$  for small  $k$ ’s. An illustration for this cooling schedule is given in Figure 3.2. In implementation,  $k$  is doubled if a stopping criterion is

met. We find the following type of relative error

$$\frac{\|\boldsymbol{\beta}^{(j+1)} - \boldsymbol{\beta}^{(j)}\|}{\|\Delta^{(j)}(k)\|}$$

makes a good criterion, where  $\Delta^{(j)}(k) \triangleq \boldsymbol{\gamma}_e^{(j)}(k) - \boldsymbol{\gamma}_o^{(j)}(k)$ . Due to (3.31), we may also use

$$k^2 \cdot \|\boldsymbol{\beta}^{(j+1)} - \boldsymbol{\beta}^{(j)}\|. \quad (3.37)$$

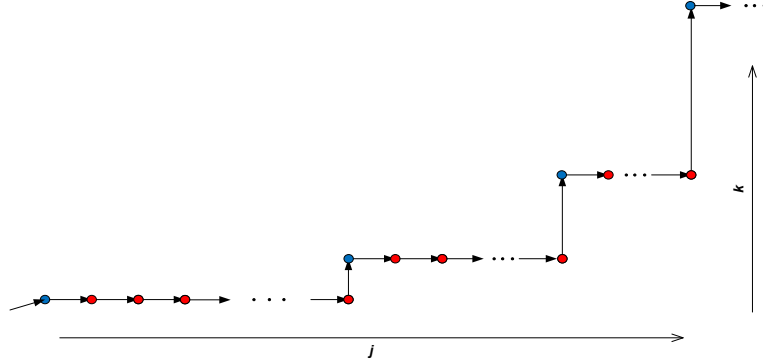


Figure 3.2: Homogenous updating in AA.

Accelerating the inner  $j$ -convergence is trickier because the iteration here is nonlinear and nonsmooth. Introduce  $\mathbf{K}$  such that

$$\mathbf{K}^T \mathbf{K} = \frac{1}{k^2} \mathbf{H}^T \boldsymbol{\Sigma} \mathbf{H} + \mathbf{U}_\perp \mathbf{U}_\perp^T,$$

where  $\mathbf{U}_\perp$  is obtained via expanding  $\mathbf{U}$  to an orthonormal  $\tilde{\mathbf{U}} = \begin{bmatrix} \mathbf{U} & \mathbf{U}_\perp \end{bmatrix}$ . Let  $\boldsymbol{\alpha} = \mathbf{H}^T \mathbf{X}^T \mathbf{y} / k^2$ . To separate the nonlinear thresholding out, the updating kernel (3.30) is represented as (see Section 3.7)

$$\boldsymbol{\xi}^{(j+1)} = (\mathbf{I} - \mathbf{K}^T \mathbf{K}) \boldsymbol{\gamma}^{(j)} + \boldsymbol{\alpha}, \boldsymbol{\gamma}^{(j+1)} = \Theta(\boldsymbol{\xi}^{(j+1)}; \frac{\lambda}{k^2}). \quad (3.38)$$

For this nonlinear process, we consider two forms of relaxation, parameterized by  $\omega$ :

$$(I) \quad \boldsymbol{\xi}^{(j+1)} = (1 - \omega) \boldsymbol{\xi}^{(j)} + \omega((\mathbf{I} - \mathbf{K}^T \mathbf{K}) \boldsymbol{\gamma}^{(j)} + \boldsymbol{\alpha}), \boldsymbol{\gamma}^{(j+1)} = \Theta(\boldsymbol{\xi}^{(j+1)}; \frac{\lambda}{k^2}), \quad (3.39)$$

$$(II) \quad \boldsymbol{\xi}^{(j+1)} = (1 - \omega)\boldsymbol{\gamma}^{(j)} + \omega((\mathbf{I} - \mathbf{K}^T \mathbf{K})\boldsymbol{\gamma}^{(j)} + \boldsymbol{\alpha}), \boldsymbol{\gamma}^{(j+1)} = \Theta(\boldsymbol{\xi}^{(j+1)}; \omega \cdot \frac{\lambda}{k^2}). \quad (3.40)$$

Both relaxations seem to converge and yield an optimal solution when  $0 < \omega < 2$ . When  $\omega = 1$ , they degenerate to the same nonrelaxation case (3.38) discussed earlier. Before proceeding, it is convenient to introduce some operators –  $T_k, \Theta_k, \tilde{T}_k, \bar{T}_k$ : for any vector  $\mathbf{v}$ ,  $T_k \circ \mathbf{v} = \mathbf{J}\mathbf{v} + \boldsymbol{\alpha}$ ,  $\forall \mathbf{v}$ , with  $\mathbf{J} = \mathbf{I} - \omega \mathbf{K}^T \mathbf{K}$ ,  $\Theta_k \circ \mathbf{v} = \Theta(\mathbf{v}; \lambda/k^2)$ ;  $\tilde{T}_k = \Theta_k \circ T_k$ ,  $\bar{T}_k = T_k \circ \Theta_k$ .

**Proposition 4** *For relaxation (II), given any  $\boldsymbol{\gamma}^{(0)}$ ,  $\boldsymbol{\gamma}^{(j)}(k)$  converges to a fixed point of  $\tilde{T}_k$  as  $j \rightarrow \infty$ , provided  $0 < \omega < 2$ . And all conclusions in Theorem 6 hold under this condition except that the last statement becomes*

$$k_0 \leq \sqrt{\frac{\omega}{2}} \frac{\sigma_{\max}(\mathbf{X})}{\sigma_{\min}(\mathbf{T})}. \quad (3.41)$$

The cooling schedule Theorem (Theorem 7) also applies for such  $\omega$ 's.

The proof is based on our generalization of Daubechies *et al.* [18].

Convergence analysis is more difficult for relaxation (I). However, our extensive experiences show that  $\boldsymbol{\gamma}^{(j)}$  converges, too, for properly chosen  $\omega$ 's. Currently, we have obtained the following result:

**Proposition 5** *For relaxation (I), given any  $\boldsymbol{\gamma}^{(0)}$ ,  $\boldsymbol{\gamma}^{(j)}(k)$  converges to a fixed point of  $\tilde{T}_k$  as  $j \rightarrow \infty$ , provided  $0 < \omega \leq 1$ . If  $2\bar{T}_k - \mathbf{I}$  is nonexpansive, the same is true for  $1 < \omega < 2$ .*

The proof presented in Section 3.7 is motivated by Browder and Petryshyn's *reasonable wanderer* [12].

A very interesting special case in relaxation (I) is  $\omega = 2$ . In this situation,  $\boldsymbol{\xi}^{(j)}$  does not converge,<sup>7</sup> but  $\boldsymbol{\gamma}^{(j)}$  converges! And the limit depends on  $\mathbf{U}_\perp \mathbf{U}_\perp^T \boldsymbol{\xi}^{(0)}$ . Consequently, if  $\mathbf{U}_\perp \mathbf{U}_\perp^T \boldsymbol{\xi}^{(0)} = \mathbf{0}$ , the limit is an optimal point (a fixed point of  $\tilde{T}_k$ ). This case catches our attention because setting  $\omega = 2$  is attractive: for a bad initial point, the relaxation can reduce the number of iterations by about 40% in comparison to  $\omega = 1$  (the non-relaxation case)!

We now state the exact procedure for the accelerated annealing (AA) algorithm of the above form. Suppose  $\mathbf{X}$ ,  $\mathbf{y}$ ,  $\lambda$ ,  $\mathbf{T}(= \mathbf{U} \mathbf{D} \mathbf{V}^T)$ , and  $\mathbf{H}$  are known. In the initialization stage, set the starting value of  $\boldsymbol{\beta}^{(cur)}$  and construct  $\boldsymbol{\gamma}^{(e)}$ . Let the initial  $k$  be the bound given

---

<sup>7</sup>It has two accumulation points.

in (3.41). We will use  $\varepsilon_{outer}$ ,  $\varepsilon_{inner,a}$ ,  $\varepsilon_{inner,b}$  as the error bounds to control the iteration (starting with  $j = 0$ ), which is specified below.

ITERATION (AA)

- $\boldsymbol{\xi}^{(new)} \leftarrow (\mathbf{I} - \frac{1}{k^2} \mathbf{H}^T \boldsymbol{\Sigma} \mathbf{H}) \mathbf{U} \mathbf{U}^T \boldsymbol{\gamma}^{(e)} + \frac{1}{k^2} \mathbf{H}^T \mathbf{X}^T \mathbf{y}$ .
- If  $j > 0$ ,  $\boldsymbol{\xi}^{(new)} \leftarrow (1 - \omega) \boldsymbol{\xi}^{(cur)} + \omega \boldsymbol{\xi}^{(new)}$ .
- $\boldsymbol{\gamma}^{(e)} \leftarrow \Theta(\boldsymbol{\xi}^{(new)}; \frac{\lambda}{k^2})$ .
- $\boldsymbol{\beta}^{(new)} \leftarrow \mathbf{H} \boldsymbol{\gamma}^{(e)}$ ,  $\boldsymbol{\gamma}^{(o)} \leftarrow \mathbf{T} \boldsymbol{\beta}^{(new)}$ .
- If  $\|\boldsymbol{\beta}^{(cur)} - \boldsymbol{\beta}^{(new)}\|_\infty < \max(\varepsilon_{inner,a}/k^2, \varepsilon_{inner,b})$ 
  - If  $\|\boldsymbol{\gamma}^{(o)} - \boldsymbol{\gamma}^{(e)}\|_\infty < \varepsilon_{outer}$ , **exit**.
  - Otherwise let  $k \leftarrow 2k$ ,  $j \leftarrow 0$ .
- $\boldsymbol{\beta}^{(cur)} \leftarrow \boldsymbol{\beta}^{(new)}$ ,  $\boldsymbol{\xi}^{(cur)} \leftarrow \boldsymbol{\xi}^{(new)}$ .
- $j \leftarrow j + 1$ ; go to the next iteration.

(Note that in the second step  $\mathbf{U}_\perp \mathbf{U}_\perp^T \boldsymbol{\xi}^{(cur)} = \mathbf{0}$  after updating  $k$  each time.) This AA algorithm solves the sparse regression for any  $\mathbf{T}$ . And it does not involve complicated operations like matrix inversion. In implementation, a pathwise algorithm with warm start is preferred, where the previous estimate associated with the old value of  $\lambda$  is used as the initial point of the procedure for the current value of  $\lambda$ .

In scientific computing, *asynchronous* iterations are often used to overcome insufficient computer capacity and computational speed. In fact, in the case of lasso with  $\mathbf{T} = \mathbf{I}$ , the asynchronous updating of (3.21) (which is in vector form) leads exactly to the component-by-component iteration given by Friedman *et al.* [27]. In AA, the first three steps in the iteration (given  $\boldsymbol{\xi}^{(cur)}$ ) can be carried out in a similar asynchronous manner, which helps in reducing the total number of iterations further (but not necessarily for the total computational time when we encounter a complex  $\mathbf{T}$  that is too large to pre-compute the matrix products in AA).

The computational cost of our algorithms is primarily due to matrix multiplication and thresholding. Although an SVD for  $\mathbf{T}$  is used, it only needs a one-time calculation. Furthermore, for some regularly patterned sparsity matrix, like the fused lasso or the clustered lasso, we are able to provide explicit analytical solutions.

Let

$$\mathbf{T}_1 = \begin{bmatrix} \mathbf{I} \\ \lambda \mathbf{F}_1 \end{bmatrix} \quad \text{with } \mathbf{F}_1 = \begin{bmatrix} 1 & -1 & & & \\ & 1 & -1 & & \\ & & \cdots & \cdots & \\ & & & 1 & -1 \end{bmatrix} \quad (3.42)$$

be the sparsity matrix for the fused lasso, and let

$$\mathbf{T}_2 = \begin{bmatrix} \mathbf{I} \\ \lambda \mathbf{F}_2 \end{bmatrix} \quad (3.43)$$

denote the sparsity matrix for the clustered lasso, where  $\mathbf{F}_2$  is a pairwise difference matrix.

It can be defined by

$$\mathbf{F}_2(i, j) = \begin{cases} 1, & \text{if } j = \alpha_i \\ -1, & \text{if } j = \beta_i \\ 0, & \text{otherwise} \end{cases} \quad (3.44)$$

for  $i = 1, \dots, d(d-1)/2$ , with  $\{(\alpha_i, \beta_i)\}$  enumerating all possible pairwise combinations of  $\{1, 2, \dots, d\}$ . Without loss of generality, assume  $\alpha_{d(d-1)/2-2} = d-2$ ,  $\beta_{d(d-1)/2-2} = d-1$ ,  $\alpha_{d(d-1)/2-1} = d-2$ ,  $\beta_{d(d-1)/2-1} = d$ ,  $\alpha_{d(d-1)/2} = d-1$ ,  $\beta_{d(d-1)/2} = d$ ; that is, the bottom right 3-by-3 submatrix of  $\mathbf{F}_2$  is  $\begin{bmatrix} 1 & -1 & \\ 1 & & -1 \\ & 1 & -1 \end{bmatrix}$ .

**Proposition 6** *The following formulas provide the SVDs for the fused lasso and the clustered lasso, with  $\mathbf{F}_1 = \mathbf{U}_1 \mathbf{D}_1 \mathbf{V}_1^T$ ,  $\mathbf{T}_1 = \tilde{\mathbf{U}}_1 \tilde{\mathbf{D}}_1 \tilde{\mathbf{V}}_1^T$ ,  $\mathbf{F}_2 = \mathbf{U}_2 \mathbf{D}_2 \mathbf{V}_2^T$ , and  $\mathbf{T}_2 = \tilde{\mathbf{U}}_2 \tilde{\mathbf{D}}_2 \tilde{\mathbf{V}}_2^T$ :*

1.  $\mathbf{U}_1 = \sqrt{\frac{2}{d}} \left[ \sin \left( \frac{ij\pi}{n} \right) \right]_{(d-1) \times (d-1)}$ ,  $\mathbf{D}_1 = \text{diag}\{2 \sin \left( \frac{i\pi}{2d} \right)\}_{(d-1) \times (d-1)}$ ,  
 $\mathbf{V}_1 = \sqrt{\frac{2}{d}} \left[ \cos \left( \frac{(2i-1)j\pi}{2n} \right) \right]_{d \times (d-1)}$ .
2.  $\tilde{\mathbf{U}}_1 = \begin{bmatrix} \frac{1}{\sqrt{d}} \mathbf{1}_{d \times 1} & \mathbf{V}_1 (\mathbf{I} + \lambda^2 \mathbf{D}_1^2)^{-\frac{1}{2}} \\ \mathbf{0}_{(d-1) \times 1} & \mathbf{U}_1 \cdot \lambda \mathbf{D}_1 (\mathbf{I} + \lambda^2 \mathbf{D}_1^2)^{-\frac{1}{2}} \end{bmatrix}$ ,  $\tilde{\mathbf{D}}_1 = \begin{bmatrix} 1_{1 \times 1} & \\ & (\mathbf{I} + \lambda^2 \mathbf{D}_1^2)^{\frac{1}{2}} \end{bmatrix}$ ,  
 $\tilde{\mathbf{V}}_1 = \begin{bmatrix} \frac{1}{\sqrt{d}} \mathbf{1}_{d \times 1} & \mathbf{V}_1 \end{bmatrix}$ .
3.  $\mathbf{U}_2 = \left[ \mathbf{u}_{21} \quad \frac{1}{\sqrt{d}} \mathbf{F}_2 \mathbf{V}_1 \right]$ ,  $\mathbf{D}_2 = \text{diag}\{0, \sqrt{d}, \dots, \sqrt{d}\}$ ,  $\mathbf{V}_2 = \tilde{\mathbf{V}}_1$ ,  $\forall d \geq 3$ ,  
 where  $\mathbf{u}_{21} = \frac{1}{\sqrt{3}} \left[ 0 \quad \cdots \quad 0 \quad 1 \quad -1 \quad 1 \right]^T$ .

$$4. \begin{aligned} \tilde{\mathbf{U}}_2 &= \begin{bmatrix} \frac{1}{\sqrt{d}} \mathbf{1}_{d \times 1} & \frac{1}{\sqrt{1+\lambda^2 d}} \mathbf{V}_1 \\ \mathbf{0} & \frac{\lambda}{\sqrt{1+\lambda^2 d}} \mathbf{F}_2 \mathbf{V}_1 \end{bmatrix}, \quad \tilde{\mathbf{D}}_2 = \text{diag}\{1, \sqrt{1+\lambda^2 d}, \dots, \sqrt{1+\lambda^2 d}\}, \\ \tilde{\mathbf{V}}_2 &= \mathbf{V}_2 = \tilde{\mathbf{V}}_1. \end{aligned}$$

To apply Proposition 6 to the DAW-CLASSO (3.17), we need to generalize our algorithms and results in this section to the weighted version of (3.22), say, given by

$$\min \frac{1}{2} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \|\boldsymbol{\Lambda}\mathbf{T}\boldsymbol{\beta}\|_1,$$

where  $\boldsymbol{\Lambda} = \text{diag}\{\lambda_i\}$  with  $\lambda_i > 0$ . This is trivial because we only need to replace the previous universal thresholding value  $\lambda$  by the componentwise thresholding values  $\lambda_i$ , then all conclusions and proofs carry over.

### 3.4.4 Results on biological data

We consider the kidney microarray data described in Section 3.2. The task of supervised clustering is quite challenging for the high-dimensional data because the size of the sparsity matrix  $\mathbf{T}$  or its left inverse  $\mathbf{H}$  can be huge ( $O(d^3)$ ) even for a medium value of  $d$ . Although we can reduce the problem size by filtering or testing — for example,  $\text{FDR} < 0.05$  gives us about 800 genes — we can only manage to run the clustered lasso with conventional convex optimization packages for  $d$  less than 110.

By contrast, in the iteration of AA, we do not really have to compute or store  $\mathbf{H}$  due to Proposition 6. In fact,  $\mathbf{U}_1$  and  $\mathbf{V}_1$  are the only dense matrices we need in calculating all the matrix-vector multiplications, and they are of order  $d \times d$ .

A very effective trick for speeding the computation of a solution path is to add predictions into the warm start: we construct the initial  $\boldsymbol{\beta}$  in the AA iteration from the **linear** extrapolation of the last two estimates.

Figure 3.3 demonstrates the gene clusters after applying DAW-CLASSO (Method B), implemented via AA, to the 800 most correlated genes. Five-fold cross-validation was used to tune the parameters. All of our efforts have paid off: the nonzero coefficients are successfully clustered. And the variable groups are directly obtained from the estimated coefficients. It seems that some of the groups might be tricky to be found by a two-step approach (modeling  $\rightarrow$  clustering, cf. Section 3.2), the clustering of which is based on the distance measure, i.e., the differences between the coefficients only. Note that in our supervised clustering approach, the clusters are optimized by the model fitting and the

fitting process automatically selects the number of clusters and the cluster size for each cluster.

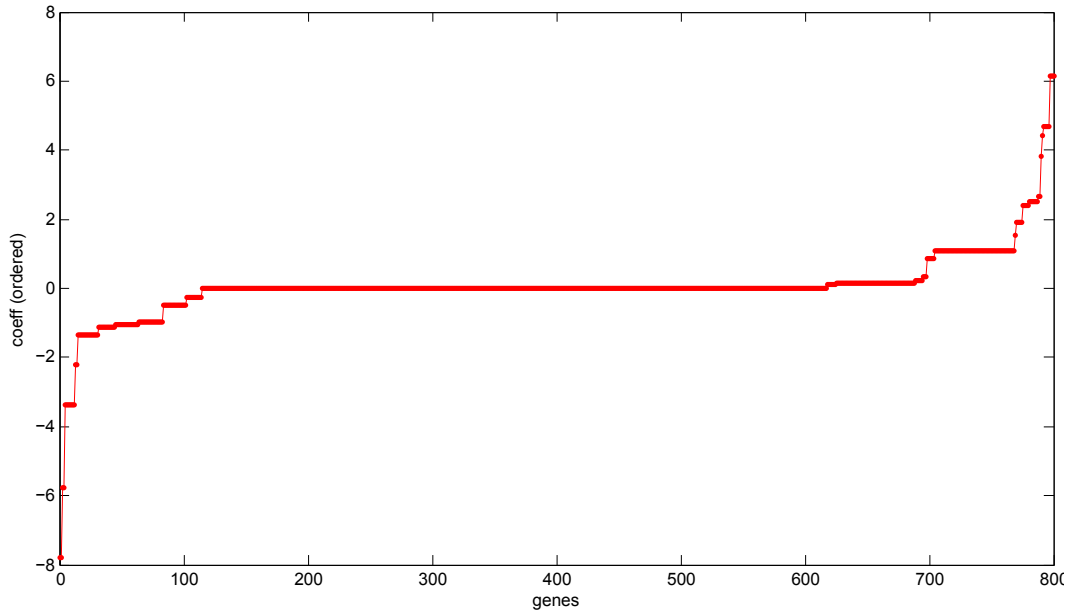


Figure 3.3: Clustered coefficients (reordered) from DAW-CLASSO on the microarray data.

### 3.5 Discussion

We studied a generic sparse regression problem with a customizable sparsity pattern matrix  $\mathbf{T}$ , motivated by, but not limited to, a supervised clustering problem in data analysis. We proposed the clustered lasso and introduced a general framework for sparse regression. Interestingly, we have found, both in practice and in theory, that the granted power of the  $l_1$ -penalty to approximate the  $l_0$ -penalty can be rather poor, say, if  $\mathbf{T}_{nz}$  is large and  $(\mathbf{T}_z, \mathbf{T}_{nz})$  is not ‘separable’. This causes serious trouble for the clustered lasso to recover all true clusterings.

In fixing the naïve  $l_1$  penalty, we noticed that Theorem 5, though general in theory, did not provide much guidance in practice for choosing the weights. We are in great need of a nonasymptotic study to tell whether a given  $\hat{\beta}_{wts}$  is accurate enough to bring benefits to the model sparsity or test error or even both.

The data augmentation technique, which can be viewed as an empirical Bayesian method,

uses Gaussian priors in our design to reduce the test error. The nondiagonal manner corresponds to a multivariate Gaussian distribution with a *degenerate* nonidentity covariance matrix, where one degree of freedom is saved for the tuning of  $\lambda$ . It is more robust to a not-so-good  $\hat{\beta}_{aug}$  than the diagonal way.

Combining weights and data-augmentation is an effective way to increase the model sparsity and reduce the test error simultaneously. In the next chapter (or see [47]), we present a totally different idea to achieve this.

Regarding the computation problem, our AA algorithm is able to handle a large  $\mathbf{T}$  and/or a large  $\mathbf{X}$  in practice, but also raises a few interesting open problems, like the accuracy of the stopping criterion (3.37) in homogenous updating, and the convergence analysis for relaxation (I) and the asynchronous updating. These studies are absolutely nontrivial due to the nonexpansive nature of the underlying operators, and the nonlinearity caused by the thresholding. Particularly, we would like to obtain some further convergence rate results of the inner/outer iteration for a singular  $\Sigma$ .

### 3.6 Proofs of Proposition 1, Proposition 2, Theorem 4, and Theorem 5

For the optimization problem

$$\min \frac{1}{2} \|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \lambda \|\mathbf{T}\beta\|_1,$$

by the KKT optimality conditions (the nonsmooth version [48]),  $\hat{\beta}$  is an optimal solution if and only if there exists a  $\widetilde{\text{sgn}}(\mathbf{T}\hat{\beta})$  such that

$$\mathbf{X}^T(\mathbf{X}\hat{\beta} - \mathbf{y}) + \lambda \mathbf{T}^T \widetilde{\text{sgn}}(\mathbf{T}\hat{\beta}) = \mathbf{0}. \quad (3.45)$$

Equivalently,

$$\hat{\beta} = \frac{1}{n} \Sigma^{-1} (\mathbf{X}^T \mathbf{y} - \lambda \mathbf{T}^T \widetilde{\text{sgn}}(\mathbf{T}\hat{\beta})),$$

or

$$\hat{\beta} = \beta + \frac{1}{n} \Sigma^{-1} \mathbf{X}^T \epsilon - \frac{\lambda}{n} \Sigma^{-1} \mathbf{T}^T \widetilde{\text{sgn}}(\mathbf{T}\hat{\beta}). \quad (3.46)$$

#### • Proof of Proposition 1

The proof is obvious by noticing that

$$\frac{1}{n}\boldsymbol{\Sigma}^{-1}\mathbf{X}^T\boldsymbol{\epsilon}\sim N(\mathbf{0},\frac{\sigma^2}{n}\boldsymbol{\Sigma}^{-1})=O_p\left(\frac{1}{\sqrt{n}}\right)=o_p(1)$$

and

$$\frac{\lambda}{n}\boldsymbol{\Sigma}^{-1}\mathbf{T}^T\widetilde{\text{sgn}}(\mathbf{T}\hat{\boldsymbol{\beta}})=\frac{\lambda}{n}O_p(1)=o_p(1). \quad \blacksquare$$

• **Proof of Proposition 2**

Assume for the moment

$$\lambda/\sqrt{n}\rightarrow\lambda_0\geq 0. \quad (3.47)$$

We first develop a  $\sqrt{n}$ -consistent result similar to Knight and Fu [34] but in a general situation:

**Lemma 3** *Under the assumptions in the Proposition 2 and (3.47),  $\sqrt{n}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})\Rightarrow\boldsymbol{\delta}_o$ , where  $\boldsymbol{\delta}_o$  is defined by*

$$\arg\min_{\boldsymbol{\delta}}\frac{1}{2}\boldsymbol{\delta}^T\mathbf{C}\boldsymbol{\delta}-\mathbf{r}^T\boldsymbol{\delta}+\lambda_0\left(\text{sgn}(\mathbf{T}_{nz}\boldsymbol{\beta})^T\mathbf{T}_{nz}\boldsymbol{\delta}+\|\mathbf{T}_z\boldsymbol{\delta}\|_1\right),$$

with  $z=\{i:(\mathbf{T}\boldsymbol{\beta})_i=0\}$ ,  $nz=\{i:(\mathbf{T}\boldsymbol{\beta})_i\neq 0\}$ , and  $\mathbf{r}\sim N(\mathbf{0},\sigma^2\mathbf{C})$ .

In fact, from the KKT equation (3.46),  $\hat{\boldsymbol{\delta}}\triangleq\sqrt{n}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})$  satisfies

$$\hat{\boldsymbol{\delta}}=\frac{1}{\sqrt{n}}\boldsymbol{\Sigma}^{-1}\mathbf{X}^T\boldsymbol{\epsilon}-\frac{\lambda}{\sqrt{n}}\boldsymbol{\Sigma}^{-1}\mathbf{T}^T\widetilde{\text{sgn}}\left(\frac{1}{\sqrt{n}}\mathbf{T}\hat{\boldsymbol{\delta}}+\mathbf{T}\boldsymbol{\beta}\right).$$

So  $\hat{\boldsymbol{\delta}}$  solves  $\frac{1}{2}\|\frac{1}{\sqrt{n}}\mathbf{X}\boldsymbol{\delta}-\boldsymbol{\epsilon}\|_2^2+\lambda\|\frac{1}{\sqrt{n}}\mathbf{T}\boldsymbol{\delta}+\mathbf{T}\boldsymbol{\beta}\|_1$ , or

$$\frac{1}{2}\|\frac{1}{\sqrt{n}}\mathbf{X}\boldsymbol{\delta}-\boldsymbol{\epsilon}\|_2^2-\frac{1}{2}\|\boldsymbol{\epsilon}\|_2^2+\lambda\|\frac{1}{\sqrt{n}}\mathbf{T}\boldsymbol{\delta}+\mathbf{T}\boldsymbol{\beta}\|_1-\lambda\|\mathbf{T}\boldsymbol{\beta}\|_1\triangleq f(\boldsymbol{\delta}).$$

Noticing that  $f(\boldsymbol{\delta})\Rightarrow g(\boldsymbol{\delta})$ ,  $\sqrt{n}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})\Rightarrow\boldsymbol{\delta}_o$  follows by Geyer [32].

We need to show  $\limsup_{n\rightarrow\infty}P(\mathbf{T}_z\hat{\boldsymbol{\beta}}=\mathbf{0})<1$ . Observing  $\{\boldsymbol{\beta}:\mathbf{T}_z\boldsymbol{\beta}=\mathbf{0}\}$  is a closed set,  $\limsup_{n\rightarrow\infty}P(\mathbf{T}_z\hat{\boldsymbol{\beta}}=\mathbf{0})\leq P(\mathbf{T}_z\boldsymbol{\delta}_o=\mathbf{0})\triangleq p_0$ .  $\boldsymbol{\delta}_o$  satisfies

$$\mathbf{C}\boldsymbol{\delta}_o-\mathbf{r}+\lambda_0\mathbf{T}_z^T\widetilde{\text{sgn}}(\mathbf{T}_z\boldsymbol{\delta}_o)+\lambda_0\mathbf{T}_{nz}^T\text{sgn}(\mathbf{T}_{nz}\boldsymbol{\beta})=\mathbf{0}.$$

Clearly,  $p_0 < 1$  if  $\lambda_0 = 0$ . Suppose  $\lambda_0 > 0$ .  $\mathbf{T}_z \boldsymbol{\delta}_o = \mathbf{0}$  means

$$\mathbf{T}_z \mathbf{C}^{-1} \mathbf{T}_z^T \widetilde{\text{sgn}}(\mathbf{T}_z \boldsymbol{\delta}_o) = \frac{1}{\lambda_0} \mathbf{T}_z \mathbf{C}^{-1} \mathbf{r} - \mathbf{T}_z \mathbf{C}^{-1} \mathbf{T}_{nz}^T \text{sgn}(\mathbf{T}_{nz} \boldsymbol{\beta}),$$

which implies

$$\begin{aligned} \mathbf{T}_z \mathbf{C}^{-1} \mathbf{T}_z^T \cdot \mathbf{s} &= \frac{1}{\lambda_0} \mathbf{T}_z \mathbf{C}^{-1} \mathbf{r} - \mathbf{T}_z \mathbf{C}^{-1} \mathbf{T}_{nz}^T \cdot \text{sgn}(\mathbf{T}_{nz} \boldsymbol{\beta}) \\ &\text{is solvable in the solution space } \{\mathbf{s} : \|\mathbf{s}\|_\infty \leq 1\}. \end{aligned} \quad (3.48)$$

**Lemma 4** *Let  $\mathbf{A}$  be a positive semi-definite matrix with the spectral decomposition given by, say,  $\mathbf{A} = \mathbf{U} \mathbf{D} \mathbf{U}^T = \sum d_i \mathbf{u}_i \mathbf{u}_i^T$ . Define  $z' = \{i : d_i = 0\}$ ,  $nz' = \{i : d_i \neq 0\}$ , and the generalized inverse  $\mathbf{A}^+ = \mathbf{U} \mathbf{D}^+ \mathbf{U}^T = \mathbf{U}_{nz'} \mathbf{D}_{nz'}^{-1} \mathbf{U}_{nz'}^T$ . Then  $\mathbf{A} \mathbf{s} = \boldsymbol{\alpha}$  if and only if (i)  $\mathbf{s} = \mathbf{A}^+ \boldsymbol{\alpha} + \mathbf{U}_{z'} \boldsymbol{\eta}$  for some  $\boldsymbol{\eta}$  and (ii)  $\mathbf{U}_{z'}^T \boldsymbol{\alpha} = \mathbf{0}$ .*

The proof is omitted.

Apply Lemma 4 to the problem (3.48), with  $\mathbf{A} = \mathbf{T}_z \mathbf{C}^{-1} \mathbf{T}_z^T$ ,  $\boldsymbol{\alpha} = \frac{1}{\lambda_0} \mathbf{T}_z \mathbf{C}^{-1} \mathbf{r} - \mathbf{T}_z \mathbf{C}^{-1} \mathbf{T}_{nz}^T \text{sgn}(\mathbf{T}_{nz} \boldsymbol{\beta})$ . (Note that condition (ii) is naturally satisfied, because

$$\mathbf{U}_{z'}^T \mathbf{A} = \mathbf{0} \Rightarrow \mathbf{U}_{z'}^T \mathbf{A} \mathbf{U}_{z'} = \mathbf{0} \Rightarrow \mathbf{U}_{z'}^T \mathbf{T}_z \mathbf{C}^{-1/2} \Rightarrow \mathbf{U}_{z'}^T \mathbf{T}_z = \mathbf{0},$$

and so  $\mathbf{U}_{z'}^T \boldsymbol{\alpha} = \mathbf{0}$ .) Then (3.48) implies  $\exists \boldsymbol{\eta}$  s.t.  $\|\mathbf{A}^+ \boldsymbol{\alpha} + \mathbf{U}_{z'} \boldsymbol{\eta}\|_\infty \leq 1$ , or

$$\left\| \begin{bmatrix} \mathbf{U}_{z'} & \mathbf{U}_{nz'} \end{bmatrix} \begin{bmatrix} \boldsymbol{\eta} \\ \mathbf{D}_{nz'}^{-1} \mathbf{U}_{nz'}^T \boldsymbol{\alpha} \end{bmatrix} \right\|_\infty \leq 1.$$

Observing that  $\begin{bmatrix} \mathbf{U}_{z'} & \mathbf{U}_{nz'} \end{bmatrix}$  is an orthonormal matrix, say, of size  $m$ -by- $m$ , we know

$$\|\mathbf{D}_{nz'}^{-1} \mathbf{U}_{nz'}^T \boldsymbol{\alpha}\|_\infty \leq \sqrt{m}.$$

Consequently, given  $\mathbf{r} \sim N(\mathbf{0}, \sigma^2 \mathbf{I})$ ,

$$p_0 \leq P(\|\mathbf{D}_{nz'}^{-1} \mathbf{U}_{nz'}^T (\frac{1}{\lambda_0} \mathbf{T}_z \mathbf{C}^{-1} \mathbf{r} - \mathbf{T}_z \mathbf{C}^{-1} \mathbf{T}_{nz}^T \text{sgn}(\mathbf{T}_{nz} \boldsymbol{\beta}))\|_\infty \leq \sqrt{m}) < 1.$$

For the general case  $\lambda = O(n)$ , if  $\hat{\boldsymbol{\beta}}$  were zero  $s$ -consistent w.r.t.  $\mathbf{T}_z$  for some sequence  $\lambda(n)$ , there must exist a subsequence  $\lambda(n_k)$  with  $\lambda(n_k)/n_k \rightarrow \lambda_0$  for some  $\lambda_0 \geq 0$ , such that

$\beta(n_k)$  is zero  $s$ -consistent w.r.t  $\mathbf{T}_z$ . This contradicts the above argument.  $\blacksquare$

• **Proof of Theorem 4**

First, it is easy to derive an asymptotic result similar to Lemma 3:

$$\frac{n}{\lambda}(\hat{\beta} - \beta) \Rightarrow \delta_o, \quad (3.49)$$

where  $\delta_o$  is nonrandom, defined by

$$\arg \min_{\delta} \frac{1}{2} \delta^T \mathbf{C} \delta + (\text{sgn}(\mathbf{T}_{nz} \beta))^T \mathbf{T}_{nz} \delta + \|\mathbf{T}_z \delta\|_1.$$

So the KKT equation for  $\delta_o$  is

$$\mathbf{C} \delta_o + \mathbf{T}_z^T \widetilde{\text{sgn}}(\mathbf{T}_z \delta_o) + \mathbf{T}_{nz}^T \text{sgn}(\mathbf{T}_{nz} \beta) = \mathbf{0}. \quad (3.50)$$

Recall that  $\hat{\beta}$  is an optimal solution if and only if (3.46) holds. Therefore,

$$\begin{aligned} \mathbf{T}_1 \hat{\beta} &= \mathbf{T}_1 \left( \frac{1}{n} \Sigma^{-1} \mathbf{X}^T \epsilon \right) - \frac{\lambda}{n} \mathbf{T}_1 \Sigma^{-1} \mathbf{T}^T \widetilde{\text{sgn}}(\mathbf{T} \hat{\beta}) \\ &= \mathbf{T}_1 \left( \frac{1}{n} \Sigma^{-1} \mathbf{X}^T \epsilon \right) - \frac{\lambda}{n} \mathbf{T}_1 \Sigma^{-1} \mathbf{T}_1^T \widetilde{\text{sgn}}(\mathbf{T}_1 \hat{\beta}) - \frac{\lambda}{n} \mathbf{T}_1 \Sigma^{-1} \mathbf{T}_2^T \widetilde{\text{sgn}}(\mathbf{T}_2 \hat{\beta}). \end{aligned}$$

Thus

$$\mathbf{T}_1 \Sigma^{-1} \mathbf{T}_1^T \widetilde{\text{sgn}}(\mathbf{T}_1 \hat{\beta}) = -\mathbf{T}_1 \Sigma^{-1} \mathbf{T}_2^T \widetilde{\text{sgn}}(\mathbf{T}_2 \hat{\beta}) + \frac{\sqrt{n}}{\lambda} \delta' - \frac{n}{\lambda} \mathbf{T}_1 \hat{\beta}, \quad (3.51)$$

with  $\delta' = \mathbf{T}_1 \Sigma^{-1} \mathbf{X}^T \epsilon / \sqrt{n} \sim N(\mathbf{0}, \mathbf{T}_1 \Sigma^{-1} \mathbf{T}_1^T)$ . Apply Lemma 4 with

$$\mathbf{A} = \mathbf{T}_1 \Sigma^{-1} \mathbf{T}_1^T, \quad \boldsymbol{\alpha} = -\mathbf{T}_1 \Sigma^{-1} \mathbf{T}_2^T \widetilde{\text{sgn}}(\mathbf{T}_2 \hat{\beta}) + \frac{\sqrt{n}}{\lambda} \delta' - \frac{n}{\lambda} \mathbf{T}_1 \hat{\beta}.$$

Again, condition (ii) is naturally satisfied because  $\mathbf{U}_z \mathbf{T}_1 = \mathbf{0}$ . So (3.51) is equivalent to  $\widetilde{\text{sgn}}(\mathbf{T}_1 \hat{\beta}) = (\mathbf{T}_1 \Sigma^{-1} \mathbf{T}_1^T)^+ \boldsymbol{\alpha} + \mathbf{U}_z \boldsymbol{\eta}$  for some  $\boldsymbol{\eta}$ . It is important to point out that even the original KKT equation (3.46) does not resolve the ambiguity of  $\boldsymbol{\eta}$  (since  $\mathbf{T}_1^T \mathbf{U}_z \boldsymbol{\eta} = \mathbf{0} \cdot \boldsymbol{\eta} = \mathbf{0}$ ). So a sufficient condition for  $\mathbf{T}_1 \hat{\beta} = \mathbf{0}$  is

$$\|(\mathbf{T}_1 \Sigma^{-1} \mathbf{T}_1^T)^+ \boldsymbol{\alpha}\|_{\infty} < 1. \quad (3.52)$$

And a necessary condition for  $\mathbf{T} \hat{\beta} = \mathbf{0}$  is  $\|(\mathbf{T}_1 \Sigma^{-1} \mathbf{T}_1^T)^+ \boldsymbol{\alpha} + \mathbf{U}_z \boldsymbol{\eta}\|_{\infty} \leq 1$  for some  $\boldsymbol{\eta}$ . It

follows that

$$\begin{aligned} \|(\mathbf{T}_1 \boldsymbol{\Sigma}^{-1} \mathbf{T}_1^T)^+ \boldsymbol{\alpha}\|_\infty &= \|\mathbf{U}_{nz'} \mathbf{U}_{nz'}^T (\mathbf{U}_{nz'} \mathbf{D}_{nz'}^{-1} \mathbf{U}_{nz'}^T \boldsymbol{\alpha} + \mathbf{U}_{z'} \boldsymbol{\eta})\|_\infty \\ &= \|\mathbf{U}_{nz'} \mathbf{U}_{nz'}^T ((\mathbf{T}_1 \boldsymbol{\Sigma}^{-1} \mathbf{T}_1^T)^+ \boldsymbol{\alpha} + \mathbf{U}_{z'} \boldsymbol{\eta})\|_\infty \\ &\leq \|\mathbf{U}_{nz'} \mathbf{U}_{nz'}^T\|_\infty = \|(\mathbf{T}_1 \boldsymbol{\Sigma}^{-1} \mathbf{T}_1^T)^+ (\mathbf{T}_1 \boldsymbol{\Sigma}^{-1} \mathbf{T}_1^T)\|_\infty. \end{aligned}$$

That is,

$$\|(\mathbf{T}_1 \boldsymbol{\Sigma}^{-1} \mathbf{T}_1^T)^+ \boldsymbol{\alpha}\|_\infty \leq \|(\mathbf{T}_1 \boldsymbol{\Sigma}^{-1} \mathbf{T}_1^T)^+ (\mathbf{T}_1 \boldsymbol{\Sigma}^{-1} \mathbf{T}_1^T)\|_\infty. \quad (3.53)$$

Now study the asymptotics.

Necessity. If  $\hat{\boldsymbol{\beta}}$  is zero  $s$ -consistent w.r.t.  $\mathbf{T}_1$ , then from (3.49),  $\mathbf{T}_1 \boldsymbol{\delta}_o = \mathbf{0}$ , and so  $\frac{n}{\lambda} \mathbf{T}_1 \hat{\boldsymbol{\beta}} \xrightarrow{P} \mathbf{0}$ . In addition,  $\frac{\sqrt{n}}{\lambda} \boldsymbol{\delta}' = o_p(1)$ . Hence

$$\|(\mathbf{T}_1 \boldsymbol{\Sigma}^{-1} \mathbf{T}_1^T)^+ \mathbf{T}_1 \boldsymbol{\Sigma}^{-1} \mathbf{T}_2^T \widetilde{\text{sgn}}(\mathbf{T}_2 \hat{\boldsymbol{\beta}})\|_\infty \leq \|(\mathbf{T}_1 \mathbf{C}^{-1} \mathbf{T}_1^T)^+ (\mathbf{T}_1 \mathbf{C}^{-1} \mathbf{T}_1^T)\|_\infty + \epsilon$$

with probability tending to 1, for any  $\epsilon > 0$ . Note  $\widetilde{\text{sgn}}(\mathbf{T}_2 \hat{\boldsymbol{\beta}})$  is bounded. There exists a subsequence indexed by  $n_k$  such that  $\widetilde{\text{sgn}}(\mathbf{T}_2 \hat{\boldsymbol{\beta}}_{n_k}) \rightarrow \mathbf{s}$  with probability 1. By Proposition 1, we immediately know  $\mathbf{s} \in \widetilde{\text{Sgn}}(\mathbf{T}_2 \boldsymbol{\beta})$ . Thus

$$\|(\mathbf{T}_1 \mathbf{C}^{-1} \mathbf{T}_1^T)^+ \mathbf{T}_1 \mathbf{C}^{-1} \mathbf{T}_2^T \mathbf{s}\|_\infty \leq \|(\mathbf{T}_1 \mathbf{C}^{-1} \mathbf{T}_1^T)^+ (\mathbf{T}_1 \mathbf{C}^{-1} \mathbf{T}_1^T)\|_\infty + \epsilon$$

with probability 1, for any  $\epsilon > 0$ . Then necessary condition follows.

Sufficiency. Our goal is to show  $P(\|(\mathbf{T}_1 \boldsymbol{\Sigma}^{-1} \mathbf{T}_1^T)^+ \boldsymbol{\alpha}\|_\infty < 1) \rightarrow 1$  given (3.10). Suppose  $\liminf_{n \rightarrow \infty} P(\|(\mathbf{T}_1 \boldsymbol{\Sigma}^{-1} \mathbf{T}_1^T)^+ \boldsymbol{\alpha}\|_\infty \geq 1) > 0$ . First, since  $\mathbf{T}_1 \subset \mathbf{T}_z$ , if we write  $\mathbf{T}_z$  as  $\begin{bmatrix} \mathbf{T}_1 \\ \mathbf{T}_{2z} \end{bmatrix}$  with  $\mathbf{T}_{2z} \subset \mathbf{T}_2$ , obviously,  $\widetilde{\text{sgn}}(\mathbf{T}_{2z} \boldsymbol{\delta}_o) \in \widetilde{\text{Sgn}}(\mathbf{T}_{2z} \boldsymbol{\beta})$ . Then, repeating the argument for (3.52), we know (3.10) is sufficient to get  $\mathbf{T}_1 \boldsymbol{\delta}_o = \mathbf{0}$  from the KKT equation (3.50).

Likewise, we can find a subsequence indexed by  $n_k$  such that  $\widetilde{\text{sgn}}(\mathbf{T}_2 \hat{\boldsymbol{\beta}}_{n_k}) \rightarrow \mathbf{s} \in \widetilde{\text{Sgn}}(\mathbf{T}_2 \boldsymbol{\beta})$ ,  $\frac{n}{\lambda} \boldsymbol{\delta}' \rightarrow \mathbf{0}$ ,  $\frac{n}{\lambda} \mathbf{T}_1 \hat{\boldsymbol{\beta}} \rightarrow \mathbf{0}$ , and  $\boldsymbol{\Sigma}_{n_k} \rightarrow \mathbf{C}$  with probability 1. So we get

$$P(\|(\mathbf{T}_1 \mathbf{C}^{-1} \mathbf{T}_1^T)^+ \mathbf{T}_1 \mathbf{C}^{-1} \mathbf{T}_2^T \mathbf{s}\|_\infty \geq 1) > 0,$$

i.e.,

$$\|(\mathbf{T}_1 \mathbf{C}^{-1} \mathbf{T}_1^T)^+ \mathbf{T}_1 \mathbf{C}^{-1} \mathbf{T}_2^T \mathbf{s}\|_\infty \geq 1,$$

which contradicts the assumption.  $\blacksquare$

• **Proof of Theorem 5**

Define  $\mathbf{W} = \text{diag}\{w_i\}$ ,  $\mathbf{W}_z = \text{diag}\{w_i\}_{j \in z}$ ,  $\mathbf{W}_{nz} = \text{diag}\{w_i\}_{j \in nz}$ . Then the weighted sparse regression (3.14) just replaces the  $\mathbf{T}$  in (3.8) by  $\mathbf{W}\mathbf{T}$ . Define  $\hat{\boldsymbol{\delta}} = a(n, \lambda)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$  for some sequence  $a(n, \lambda)$ . Similar to the derivation Lemma 3,  $\hat{\boldsymbol{\delta}}$  solves

$$\min \frac{1}{2} \boldsymbol{\delta}^T \boldsymbol{\Sigma} \boldsymbol{\delta} - \frac{a}{\sqrt{n}} \left( \frac{1}{\sqrt{n}} \mathbf{x}^T \boldsymbol{\epsilon} \right)^T \boldsymbol{\delta} + \frac{\lambda a}{n} \|\mathbf{W}_z \mathbf{T}_z \boldsymbol{\delta}\|_1 + \frac{\lambda a}{n} \text{sgn}(\mathbf{T}_{nz} \boldsymbol{\beta})^T \mathbf{W}_{nz} \mathbf{T}_{nz} \boldsymbol{\delta}.$$

Following the lines of [61], one can prove that if (i)  $\lim \frac{a}{\sqrt{n}}$  exist (say equal to  $a_0$ ), (ii)  $\frac{n}{\lambda a} A \rightarrow 0$ , and (iii)  $\frac{\lambda a}{n} B \rightarrow 0$ , then

$$\hat{\boldsymbol{\delta}} = a(n, \lambda)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \Rightarrow \arg \left( \min_{\boldsymbol{\delta}} \frac{1}{2} \boldsymbol{\delta}^T \mathbf{C} \boldsymbol{\delta} - a_0 \mathbf{r}^T \boldsymbol{\delta}, \text{ s.t. } \mathbf{T}_z \boldsymbol{\delta} = \mathbf{0} \right), \quad (3.54)$$

with  $\mathbf{r} \sim N(\mathbf{0}, \sigma^2 \mathbf{C})$ . To make sure such  $a(n, \lambda)$  exists, it is enough to have:  $\frac{nA}{\lambda} \ll \frac{n}{\lambda B}$  and  $\frac{nA}{\lambda} \ll \sqrt{n}$  (with  $P \ll Q$  meaning  $\lim P/Q \rightarrow 0$ ). That is, if

$$\frac{\sqrt{n}}{\lambda} A(n) \rightarrow 0, A(n)B(n) \rightarrow 0, \quad (3.55)$$

then  $\hat{\boldsymbol{\beta}}$  is  $a(n, \lambda)$ -consistent, for any  $a$  satisfying (i), (ii), & (iii).

On the other hand, substituting  $\mathbf{W}_z \mathbf{T}_z$  for  $\mathbf{T}_1$  and  $\mathbf{W}_{nz} \mathbf{T}_{nz}$  for  $\mathbf{T}_2$  in (3.52), we obtain a sufficient condition for  $\mathbf{T}_z \hat{\boldsymbol{\beta}} = \mathbf{0}$ :

$$\left\| \mathbf{W}_z^{-1} (\mathbf{T}_z \boldsymbol{\Sigma}^{-1} \mathbf{T}_z^T)^+ \left( -\mathbf{T}_z \boldsymbol{\Sigma}^{-1} \mathbf{T}_{nz}^T \mathbf{W}_{nz} \widehat{\text{sgn}}(\mathbf{T}_{nz} \hat{\boldsymbol{\beta}}) + \frac{\sqrt{n}}{\lambda} \frac{\mathbf{T}_z \boldsymbol{\Sigma}^{-1} \mathbf{x}^T \boldsymbol{\epsilon}}{\sqrt{n}} - \frac{n}{\lambda a(n, \lambda)} a(n, \lambda) \mathbf{T}_z \hat{\boldsymbol{\beta}} \right) \right\|_{\infty} < 1.$$

Clearly, by (3.54), (3.55), and (ii), this holds with probability tending to 1.

For the special case  $a = \sqrt{n}$ , it suffices to show  $\lambda$  satisfying  $\sqrt{n}A/\lambda \rightarrow 0$ ,  $\lambda B/\sqrt{n} \rightarrow 0$  exists. And  $\lambda = \sqrt{nA/B}$  is one possible choice.  $\blacksquare$

### 3.7 Proofs of Theorem 6, Proposition 3, Theorem 7, Proposition 4, Proposition 5, and Proposition 6

#### • Some Basic Facts

Before our formal proofs, let's state some basic facts. Recall that  $\mathbf{T} = \mathbf{U}\mathbf{D}\mathbf{V}^T$ ,  $\mathbf{H} = \mathbf{V}\mathbf{D}^{-1}\mathbf{U}^T$ , and  $\mathbf{I} - \mathbf{T}\mathbf{H} = \mathbf{U}_\perp\mathbf{U}_\perp^T$  with  $\mathbf{U}_\perp$  via expanding  $\mathbf{U}$  to get  $\tilde{\mathbf{U}} = \begin{bmatrix} \mathbf{U} & \mathbf{U}_\perp^T \end{bmatrix}$  orthonormal;  $C$  is widely used to denote a positive constant, but not necessarily the same even in a single formula. The subscripts of  $\gamma_e^{(j)}(k)$  and  $\gamma_e(k)$  are omitted for short.

From (3.30),  $\gamma_e(k)$ , or  $\gamma(k)$ , satisfies

$$\gamma(k) + \frac{\lambda}{k^2} \widehat{\text{sgn}}(\gamma(k)) = \mathbf{U}\mathbf{U}^T \gamma(k) + \frac{1}{k^2} (\mathbf{H}^T \mathbf{X}^T \mathbf{y} - \mathbf{H}^T \Sigma \mathbf{H} \gamma(k)),$$

i.e.,

$$\begin{aligned} \gamma(k) &= \frac{1}{k} \arg \min \frac{\lambda}{k} \|\gamma\|_1 + \frac{1}{2} \|\mathbf{y} - (\mathbf{X}\mathbf{H}/k) \cdot \gamma\|_2^2 + \frac{1}{2} \|\mathbf{U}_\perp \mathbf{U}_\perp^T \gamma\|_2^2 \\ \implies \gamma(k) &= \arg \min \lambda \|\gamma\|_1 + \frac{1}{2} \|\mathbf{y} - \mathbf{X}\mathbf{H} \cdot \gamma\|_2^2 + \frac{k^2}{2} \|\mathbf{U}_\perp \mathbf{U}_\perp^T \gamma\|_2^2 \end{aligned} \quad (3.56)$$

Let

$$f(\gamma) = \frac{1}{2} \|\mathbf{y} - \mathbf{X}\mathbf{H} \cdot \gamma\|_2^2 + \lambda \|\gamma\|_1, \quad (3.57)$$

$$F_k(\gamma) = f(\gamma) + \frac{k^2}{2} \|\mathbf{U}_\perp \mathbf{U}_\perp^T \gamma\|_2^2, \quad (3.58)$$

$$\Phi_k(\gamma) = \frac{1}{k^2} F_k(\gamma). \quad (3.59)$$

*Fact 1)* For any  $k$ ,  $\gamma^{(j)}(k)$  ( $j = 0, 1, \dots$ ) defined by (3.30) is the sequence of iterates solving the lasso problem  $\min_{\gamma} \Phi_k(\gamma)$ , in the way of (3.21).

This gives another important explanation of our approach from the *penalty functions*.

And we immediately know (see, e.g., [6])

*Fact 2)*  $f(\gamma(k)) \uparrow$ ,  $f(\gamma(k)) \leq f_{opt}$ .

From Fact 2),  $\lambda \|\gamma(k)\|_1 \leq f_{opt}$ . So

*Fact 3)*  $\|\gamma(k)\|$  is uniformly bounded.

The KKT equation yields

$$\mathbf{U}_\perp \mathbf{U}_\perp^T \boldsymbol{\gamma}(k) = \frac{1}{k^2} (\mathbf{H}^T \mathbf{X}^T \mathbf{y} - \mathbf{H}^T \boldsymbol{\Sigma} \mathbf{H} \boldsymbol{\gamma}(k) - \lambda \widetilde{\text{sgn}}(\boldsymbol{\gamma}(k))).$$

It follows from Fact 3) that

*Fact 4)*  $\|\Delta(k)\| = \|\mathbf{U}_\perp \mathbf{U}_\perp^T \boldsymbol{\gamma}(k)\| = O(\frac{1}{k^2})$  and  $\|\Delta(k)\|_2 \downarrow 0$ .

The latter result is due to the penalty function again.

• **Generalization of Daubechies *et al.*'s Convergence Theorem**

Although we have Fact 1), Daubechies *et al.*'s convergence theorem [18], which makes use of Opial's conditions [41] in studying the nonexpansive operators, can *not* be directly applied, because the 2-norm of the operator  $\mathbf{K}$  satisfying

$$\mathbf{K}^* \mathbf{K} = \frac{1}{k^2} \mathbf{H}^T \boldsymbol{\Sigma} \mathbf{H} + \mathbf{U}_\perp \mathbf{U}_\perp^T = \begin{bmatrix} \mathbf{U} & \mathbf{U}_\perp \end{bmatrix} \begin{bmatrix} \frac{1}{k^2} \mathbf{A} & \\ & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{U}^T \\ \mathbf{U}_\perp^T \end{bmatrix} \quad (3.60)$$

with

$$\mathbf{A} \triangleq \mathbf{D}^{-1} \mathbf{V}^T \boldsymbol{\Sigma} \mathbf{V} \mathbf{D}^{-1}, \quad (3.61)$$

is exactly 1, whereas Theorem 3.1 [18] requires  $\|\mathbf{K}\|_2 < 1$ . Therefore, we need a generalization of Daubechies *et al.*'s (weak) convergence result.

In fact, we can generalize Theorem 3.1 (or Proposition 3.11, to be more exact) in [18] to  $\mathbf{K}$  satisfying

$$\|\mathbf{K}\|_2 < \sqrt{2}, \quad (3.62)$$

which may also validate the over-relaxation technique used to speed the convergence. In this part, we will use some notations compatible with [18], with mild changes. (Note that our thresholding operator  $\Theta(\cdot; \lambda)$  uses a threshold value  $\lambda$  instead of  $\lambda/2$ .)

Let

$$\Phi(\mathbf{f}) = \frac{1}{2} \|\mathbf{K} \mathbf{f} - \mathbf{g}\|_2^2 + \lambda \|\mathbf{f}\|_1, \Phi^{\text{SUR}}(\mathbf{f}; \mathbf{a}) = \Phi(\mathbf{f}) + \frac{1}{2} (\mathbf{f} - \mathbf{a})^T \mathbf{J} (\mathbf{f} - \mathbf{a}), \quad (3.63)$$

and  $\mathbf{J} \triangleq \mathbf{I} - \mathbf{K}^* \mathbf{K}$ . The iterative process can be represented as

$$\mathbf{f}^{n+1} = \Theta(\mathbf{J} \mathbf{f}^n + \mathbf{K}^* \mathbf{g}; \lambda) \quad (3.64)$$

Since  $\|\mathbf{K}\|_2 < \sqrt{2}$ ,  $-1 < \text{eig}(\mathbf{J}) \leq 1$ , where  $\text{eig}(\mathbf{J})$  denotes any eigenvalue of  $\mathbf{J}$ . Note

that  $\Phi^{\text{SUR}}$  is still strictly convex in  $\mathbf{f}$  and Proposition 2.1 [18] holds; in particular, for  $\mathbf{f}_{opt} = \arg \min_{\mathbf{f}} \Phi^{\text{SUR}}(\mathbf{f}; \mathbf{a})$  given  $\mathbf{a}$ ,

$$\Phi^{\text{SUR}}(\mathbf{f}_{opt} + \mathbf{h}; \mathbf{a}) \geq \Phi^{\text{SUR}}(\mathbf{f}_{opt}; \mathbf{a}) + \|\mathbf{h}\|_2^2, \quad \forall \mathbf{h}. \quad (3.65)$$

Let  $\mathbf{f}^{n+1} = \arg \min_{\mathbf{f}} \Phi^{\text{SUR}}(\mathbf{f}; \mathbf{f}^n)$ . Then it is easy to get

$$\begin{aligned} & \Phi(\mathbf{f}^{n+1}) + \frac{1}{2}(\mathbf{f}^{n+1} - \mathbf{f}^n)^T \mathbf{J}(\mathbf{f}^{n+1} - \mathbf{f}^n) = \Phi^{\text{SUR}}(\mathbf{f}^{n+1}; \mathbf{f}^n) \\ & \leq \Phi^{\text{SUR}}(\mathbf{f}^n; \mathbf{f}^n) - \frac{1}{2}\|\mathbf{f}^{n+1} - \mathbf{f}^n\|_2^2 = \Phi(\mathbf{f}^n) - \frac{1}{2}\|\mathbf{f}^{n+1} - \mathbf{f}^n\|_2^2 \\ & \implies \Phi(\mathbf{f}^{n+1}) + \frac{1}{2}(\mathbf{f}^{n+1} - \mathbf{f}^n)^T (\mathbf{I} + \mathbf{J})(\mathbf{f}^{n+1} - \mathbf{f}^n) \leq \Phi(\mathbf{f}^n). \end{aligned}$$

Hence  $\Phi(\mathbf{f}^n) \downarrow$  and the series  $\sum_{n=0}^{\infty} (\mathbf{f}^{n+1} - \mathbf{f}^n)^T (\mathbf{I} + \mathbf{J})(\mathbf{f}^{n+1} - \mathbf{f}^n)$  is convergent. On the other hand, since  $\text{eig}(\mathbf{J}) > -1$ ,  $\|\mathbf{f}^{n+1} - \mathbf{f}^n\|_2 \leq A \cdot \|(\mathbf{I} + \mathbf{J})^{1/2}(\mathbf{f}^{n+1} - \mathbf{f}^n)\|_2$ , where  $A$  is some strictly positive constant.

With these facts, it is not difficult to write out the full proof for the (weak) convergence of  $\mathbf{f}^n$  for any  $\mathbf{K}$  satisfying (3.62), by making corresponding changes in Lemma 3.5 and Lemma 3.7 [18]. The details are left to the readers.  $\blacksquare$

#### • Proofs of Theorem 6 and Proposition 4

Now, with Fact 1) and the above generalization,  $\gamma^{(j)}, \gamma_o^{(j)}$  as defined by (3.30) must converge given any initial value, because  $\|\mathbf{K}\|_2 = 1 < \sqrt{2}$ . (Recall that  $\mathbf{K}$  is defined by (3.60).)

By Fact 3),  $\{\gamma(k)\}$  has at least one accumulation point. Consider a subsequence  $\gamma(k_l) \rightarrow \gamma_o$  as  $l \rightarrow \infty$ . Then  $f(\gamma_o) = \lim_{l \rightarrow \infty} f(\gamma(k_l)) \leq f_{opt}$  due to Fact 2). So any accumulation point of  $\gamma(k)$  is an optimal solution.

The convergence rate of  $\|\Delta(k)\|$  is covered by Fact 4).

From Fact 2),  $f(\gamma(k)) \uparrow$ ,  $f_{opt} - f(\gamma(k)) \geq 0$ . So  $f(\gamma(k))$  converges. Note that

$$f(\gamma) \triangleq \frac{1}{2}\|\mathbf{y} - \mathbf{X}\mathbf{H}\gamma\|_2^2 + \lambda\|\gamma\|_1 \geq \frac{1}{2}\|\mathbf{y} - \mathbf{X}\mathbf{H}\gamma\|_2^2 + \lambda\|\mathbf{U}\mathbf{U}^T\gamma\|_1 - \lambda\|\mathbf{U}_{\perp}\mathbf{U}_{\perp}^T\gamma\|_1.$$

It follows from Fact 4) that

$$f(\gamma(k)) \geq \frac{1}{2}\|\mathbf{y} - \mathbf{X}\mathbf{H}\gamma(k)\|_2^2 + \lambda\|\mathbf{U}\mathbf{U}^T\gamma(k)\|_1 - \frac{1}{k^2}C.$$

Let  $g_{opt}$  be the optimal value of

$$\min g(\gamma) \triangleq \frac{1}{2} \|\mathbf{y} - \mathbf{X}\mathbf{H}\mathbf{U}\mathbf{U}^T\gamma\|_2^2 + \lambda \|\mathbf{U}\mathbf{U}^T\gamma\|_1 \text{ s.t. } \|\mathbf{U}_\perp\mathbf{U}_\perp^T\gamma\| \leq \|\mathbf{U}_\perp\mathbf{U}_\perp^T\gamma(k)\|. \quad (3.66)$$

Then

$$f(\gamma(k)) \geq g_{opt} - \frac{C}{k^2}. \quad (3.67)$$

Observe that for any  $\gamma$  minimizing (3.66),  $\mathbf{U}\mathbf{U}^T\gamma + \theta\mathbf{U}_\perp\mathbf{U}_\perp^T\gamma$  is an optimal solution, too, for  $\forall \theta : 0 \leq \theta \leq 1$ . So it is enough to consider

$$\min g(\gamma) \text{ s.t. } \|\mathbf{U}_\perp\mathbf{U}_\perp^T\gamma\| = 0,$$

which is equivalent to

$$\min f(\gamma) \text{ s.t. } \mathbf{U}_\perp\mathbf{U}_\perp^T\gamma = \mathbf{0}$$

Thus  $\gamma_{opt}$  is always one optimal solution to (3.66) given any  $k$ . By (3.67),

$$f(\gamma(k)) \geq g(\gamma_{opt}) - \frac{C}{k^2} = f(\gamma_{opt}) - \frac{C}{k^2}.$$

For the relaxation in the form of (II), it is of the same form as (3.64) if we let

$$\mathbf{J}(\text{or } \mathbf{J}_k) = \mathbf{I} - \omega \mathbf{K}^T \mathbf{K}, \text{ for } 0 < \omega < 2, \quad (3.68)$$

with  $\omega = 1$  corresponding to the non-relaxed version (3.30) (or (3.38)). Since  $\sqrt{\omega} \cdot \|\mathbf{K}\|_2 < \sqrt{2}$ ,  $\gamma^{(j)}$  defined by (3.40) converges. Clearly, the above conclusions and proofs go through.

For the choice of  $k_0$ , our generalization guarantees the convergence if (3.62) is satisfied, and we know

$$\omega \|\mathbf{A}\|_2 \cdot \frac{1}{k^2} < 2 \iff k^2 > \frac{\omega}{2} \|\mathbf{D}^{-1}\mathbf{V}^T\mathbf{\Sigma}\mathbf{V}\mathbf{D}^{-1}\|_2.$$

Since

$$\|\mathbf{D}^{-1}\mathbf{V}^T\mathbf{\Sigma}\mathbf{V}\mathbf{D}^{-1}\|_2 \leq \frac{\sigma_{\max}^2(\mathbf{X})}{\sigma_{\min}^2(\mathbf{T})},$$

it is sufficient to let

$$k > \sqrt{\frac{\omega}{2}} \cdot \frac{\sigma_{\max}(\mathbf{X})}{\sigma_{\min}(\mathbf{T})}.$$

Hence  $k_0 \leq \sqrt{\frac{\omega}{2}} \cdot \frac{\sigma_{\max}(\mathbf{X})}{\sigma_{\min}(\mathbf{T})}$ .

The cooling schedule part of Proposition 4 is left to the proof of Theorem 7.  $\blacksquare$

• **Proof of Proposition 3**

Represent the iteration of  $\gamma^{(j)}$  by nonexpansive operators  $\tilde{T}_k, \Theta_k, T_k$ :

$$\gamma^{(j+1)} = \tilde{T}_k \circ \gamma^{(j)} = \Theta_k \circ (T_k \circ \gamma^{(j)}), \quad (3.69)$$

where  $\Theta_k \circ \mathbf{v} = \Theta(\mathbf{v}; \lambda/k^2)$ ,  $T_k \circ \mathbf{v} = \mathbf{J}_k \mathbf{v} + \boldsymbol{\alpha}_k$  with  $\boldsymbol{\alpha}_k = \mathbf{H}^T \mathbf{X}^T \mathbf{y}/k^2$ .  $\Theta_k, T_k, \tilde{T}_k$  are nonexpansive in that

$$\|\Theta_k \circ \mathbf{v} - \Theta_k \circ \mathbf{v}'\| \leq \|\mathbf{v} - \mathbf{v}'\|, \quad \|\tilde{T}_k \circ \mathbf{v} - \tilde{T}_k \circ \mathbf{v}'\| \leq \|T_k \circ \mathbf{v} - T_k \circ \mathbf{v}'\| \leq \|\mathbf{v} - \mathbf{v}'\|.$$

(See Lemma 2.2 and Lemma 3.4 of [18].) For later use, we also define  $\bar{T}_k = T_k \circ \Theta_k$ .

If  $\boldsymbol{\Sigma}$  is nonsingular,  $\text{eig}(\mathbf{K}) > 0$ , and thus  $\tilde{T}_k$  becomes a contraction. It is not difficult to show (3.34) since  $\lambda_{\min}(\mathbf{A}) = \lambda_{\min}^+ \left( \tilde{\mathbf{U}} \begin{bmatrix} \mathbf{A} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \tilde{\mathbf{U}}^T \right) = \lambda_{\min}^+(\mathbf{H}^T \boldsymbol{\Sigma} \mathbf{H})$ . And  $f(\boldsymbol{\gamma})$  is strictly convex, so  $\boldsymbol{\gamma}_{opt}$  is of course unique.

Next, we prove the finite- $k$  sign consistency result.<sup>8</sup> Introduce an important fact:

*Fact 5)* Let  $\mathbf{v}_o$  be the unique optimal solution of the convex optimization  $\min f_0(\mathbf{v}) \triangleq h(\mathbf{v}) + \|\mathbf{B}(\mathbf{v})\|_1$  with  $h, \mathbf{B}$  smooth. Define the index sets  $z = \{i : (\mathbf{B}(\mathbf{v}_o))_i = 0\}$ , and  $nz = \{i : (\mathbf{B}(\mathbf{v}_o))_i \neq 0\}$ . Let  $\mathbf{v}_{oo}$  be the optimal solution of

$$\min_{\mathbf{v}} h(\mathbf{v}) + \text{sgn}(\mathbf{B}(\mathbf{v}_o)_{nz})^T \mathbf{B}(\mathbf{v})_{nz} \text{ s.t. } (\mathbf{B}(\mathbf{v}))_z = \mathbf{0}. \quad (3.70)$$

Then  $\mathbf{v}_o = \mathbf{v}_{oo}$ .

In fact, by the generalized KKT (see, e.g., [48]),  $\mathbf{v}_o$  solves  $\min f_0(\mathbf{v})$  if and only if

$$\nabla h(\mathbf{v}_o) + D\mathbf{B}(\mathbf{v}_o)^T \widetilde{\text{sgn}}(\mathbf{B}(\mathbf{v}_o)) = \mathbf{0}.$$

Let  $\mathbf{b} = \widetilde{\text{sgn}}(\mathbf{B}(\mathbf{v}_o))$ . Then  $\min f_0(\mathbf{v}) \iff \min f_1(\mathbf{v}) \triangleq h(\mathbf{v}) + \mathbf{b}^T \mathbf{B}(\mathbf{v})$ . And we know  $b_i =$

<sup>8</sup>This proof is partly motivated by Boot's sensitivity analysis in QP [9].

$\pm 1, \forall i \in nz, b_i \in [-1, 1], \forall i \in z$ . Now consider  $\min f_2(\mathbf{v}) \triangleq h(\mathbf{v}) + \mathbf{b}_{nz}^T \cdot (\mathbf{B}(\mathbf{v}))_{nz}$  s.t.  $(\mathbf{B}(\mathbf{v}))_z = \mathbf{0}$  with an optimal solution  $\mathbf{v}_{oo}$ . We have  $f_1(\mathbf{v}_o) = f_2(\mathbf{v}_o) \geq f_2(\mathbf{v}_{oo}) = f_1(\mathbf{v}_{oo})$ . Hence  $\mathbf{v}_o = \mathbf{v}_{oo}$ .

In our problem, observe that  $\boldsymbol{\eta}_k \triangleq \mathbf{U}^T \boldsymbol{\gamma}(k), \boldsymbol{\eta}_{opt} \triangleq \mathbf{U}^T \boldsymbol{\gamma}_{opt}$  respectively solve

$$\min_{\boldsymbol{\eta}} \frac{1}{2} \|\mathbf{XVD}^{-1} \cdot \boldsymbol{\eta} - \mathbf{y}\|_2^2 + \lambda \|\mathbf{U}\boldsymbol{\eta} + \mathbf{U}_{\perp} \mathbf{U}_{\perp}^T \boldsymbol{\gamma}(k)\|_1,$$

and

$$\min \frac{1}{2} \|\mathbf{XVD}^{-1} \cdot \boldsymbol{\eta} - \mathbf{y}\|_2^2 + \lambda \|\mathbf{U}\boldsymbol{\eta}\|_1.$$

Define index sets  $z = \{i : (\boldsymbol{\gamma}_{opt})_i = 0\}$ ,  $nz = \{i : (\boldsymbol{\gamma}_{opt})_i \neq 0\}$ . In the remainder of this proof, given any index set  $I$ , we use  $\mathbf{U}_I$  to denote the submatrix of  $\mathbf{U}$  composed of its corresponding rows such that  $(\mathbf{U}\boldsymbol{\alpha})_I = \mathbf{U}_I \cdot \boldsymbol{\alpha}, \forall \boldsymbol{\alpha}$ .

Fact 5) tells us that  $\boldsymbol{\eta}_{opt}$  solves

$$\min \frac{1}{2} \|\mathbf{XVD}^{-1} \cdot \boldsymbol{\eta} - \mathbf{y}\|_2^2 + \lambda \text{sgn}((\boldsymbol{\gamma}_{opt})_{nz})^T \cdot (\mathbf{U}\boldsymbol{\eta})_{nz} \text{ s.t. } (\mathbf{U}\boldsymbol{\eta})_z = \mathbf{0}, \quad (3.71)$$

because  $\mathbf{U}\boldsymbol{\eta}_{opt} = \boldsymbol{\gamma}_{opt}$ . Clearly,  $\text{sgn}((\mathbf{U}\boldsymbol{\eta}_k)_{nz}) = \text{sgn}((\boldsymbol{\gamma}_{opt})_{nz})$  for  $k$  large enough since  $\mathbf{U}\boldsymbol{\eta}_k \rightarrow \boldsymbol{\gamma}_{opt}$ . We claim that  $(\mathbf{U}\boldsymbol{\eta}_k)_z = (\boldsymbol{\gamma}_{opt})_z = \mathbf{0}$  is also true for *any*  $k$  large enough.

Otherwise, noticing  $\boldsymbol{\gamma}_{opt}$  is finite dimensional, there must exist some index sets  $nzz \subset z$ , and  $zz = z \setminus nzz$  such that each component of  $(\mathbf{U}\boldsymbol{\eta}_{k_j})_{nzz}$  is nonzero, and  $(\mathbf{U}\boldsymbol{\eta}_{k_j})_{zz} = \mathbf{0}$ , for some subsequence  $\boldsymbol{\eta}_{k_j}$  with  $k_j \rightarrow \infty$  as  $j \rightarrow \infty$ , which implies  $\mathbf{U}_{\perp} \mathbf{U}_{\perp}^T \boldsymbol{\gamma}(k_j) \rightarrow \mathbf{0}$ . It follows that a further subsequence of  $\boldsymbol{\eta}_{k_j}$  ( $j = 0, 1, \dots$ ) asymptotically solves

$$\min \frac{1}{2} \|\mathbf{XVD}^{-1} \cdot \boldsymbol{\eta} - \mathbf{y}\|_2^2 + \lambda \text{sgn}((\boldsymbol{\gamma}_{opt})_{nz})^T \cdot (\mathbf{U}\boldsymbol{\eta})_{nz} + \lambda \mathbf{b}_{nzz}^T \cdot (\mathbf{U}_{nzz} \boldsymbol{\eta}) \text{ s.t. } \mathbf{U}_{zz} \boldsymbol{\eta} = \mathbf{0} \quad (3.72)$$

for some sign vector  $\mathbf{b}_{nzz}$  (with each component  $\pm 1$ ). Obviously, none of the rows of  $\mathbf{U}_{nzz}$  lies in the (row) space spanned by the row vectors of  $\mathbf{U}_{zz}$ . Excluding the case of degeneracy, the optimization problem does not have the same optimal solution  $\boldsymbol{\eta}_{opt}$  as (3.71). Hence the finite- $k$  sign consistency holds; and we also obtain for  $k$  large,  $\boldsymbol{\eta}_k$  solves

$$\min \frac{1}{2} \|\mathbf{XVD}^{-1} \cdot \boldsymbol{\eta} - \mathbf{y}\|_2^2 + \lambda \text{sgn}((\boldsymbol{\gamma}_{opt})_{nz})^T \cdot (\mathbf{U}\boldsymbol{\eta})_{nz} \text{ s.t. } (\mathbf{U}\boldsymbol{\eta})_z = (\mathbf{U}_{\perp} \mathbf{U}_{\perp}^T \boldsymbol{\gamma}(k))_z. \quad (3.73)$$

Note that (3.73) is a simple quadratic programming (QP) problem.

Let  $rz \subset z$  be one index set such that  $\mathbf{U}_{rz}$  has full row rank and  $\text{rank}(\mathbf{U}_{rz}) = \text{rank}(\mathbf{U}_z)$ . Since  $\boldsymbol{\eta}_k$  always exists, the optimization problem (3.73) can be simplified into

$$\min \frac{1}{2} \|\mathbf{XVD}^{-1} \cdot \boldsymbol{\eta} - \mathbf{y}\|_2^2 + \lambda \text{sgn}((\boldsymbol{\gamma}_{opt})_{nz})^T \cdot (\mathbf{U}_{nz}\boldsymbol{\eta}) \text{ s.t. } \mathbf{U}_{rz}\boldsymbol{\eta} = (\mathbf{U}_\perp \mathbf{U}_\perp^T)_{rz} \boldsymbol{\gamma}(k),$$

or

$$\min \frac{1}{2} \boldsymbol{\eta}^T \mathbf{A} \boldsymbol{\eta} - \boldsymbol{\alpha}^T \boldsymbol{\eta} \text{ s.t. } \mathbf{U}_{rz}\boldsymbol{\eta} = \boldsymbol{\delta}_k, \quad (3.74)$$

where  $\boldsymbol{\alpha} = \mathbf{D}^{-1} \mathbf{V}^T \mathbf{X}^T \mathbf{y} - \lambda \mathbf{U}^T \text{sgn}((\boldsymbol{\gamma}_{opt})_{nz})$ ,  $\boldsymbol{\delta}_k = (\mathbf{U}_\perp \mathbf{U}_\perp^T)_{rz} \boldsymbol{\gamma}(k)$ .

Solving this QP, we obtain

$$\boldsymbol{\eta}_k = \{ \mathbf{A}^{-1} \boldsymbol{\alpha} - \mathbf{A}^{-1} \mathbf{U}_{rz}^T (\mathbf{U}_{rz} \mathbf{A}^{-1} \mathbf{U}_{rz}^T)^{-1} \mathbf{U}_{rz} \mathbf{A}^{-1} \boldsymbol{\alpha} \} + \mathbf{A}^{-1} \mathbf{U}_{rz}^T (\mathbf{U}_{rz} \mathbf{A}^{-1} \mathbf{U}_{rz}^T)^{-1} \cdot \boldsymbol{\delta}_k. \quad (3.75)$$

Note that since  $\mathbf{U}_{rz}$  has full row rank,  $(\mathbf{U}_{rz} \mathbf{A}^{-1} \mathbf{U}_{rz}^T)^{-1}$  exists. Now it follows immediately that

**Lemma 5**  $\|\mathbf{U} \mathbf{U}^T \cdot (\boldsymbol{\gamma}(k) - \boldsymbol{\gamma}(k'))\| \leq C \cdot \|\mathbf{U}_\perp \mathbf{U}_\perp^T \cdot (\boldsymbol{\gamma}(k) - \boldsymbol{\gamma}(k'))\|, \quad \forall k, k'.$

Letting  $k \rightarrow \infty$ , we get the convergence rate of  $\boldsymbol{\gamma}(k)$ :  $\|\boldsymbol{\gamma}(k) - \boldsymbol{\gamma}_{opt}\| = O(1/k^2)$ . ■

• **Proof of Theorem 7**

We prove the theorem for the general relaxed case in the form of (II), where  $0 < \omega < 2$ , with  $\omega = 1$  corresponding to the non-relaxed version; see (3.68). The operators introduced in (3.69) will be used for simplicity except that  $\Theta_k$  is redefined to include  $\omega$ :  $\Theta_k \circ \mathbf{v} = \Theta(\mathbf{v}; \omega \lambda / k^2), \forall \mathbf{v}$ .

First, from

$$\begin{aligned} & \tilde{T}_{k(n)} \circ \cdots \circ \tilde{T}_{k(1)} \circ \boldsymbol{\gamma}^{(0)} - \boldsymbol{\gamma}_{opt} \\ = & \left( \tilde{T}_{k(n)} \circ \cdots \circ \tilde{T}_{k(N)} \circ (\tilde{T}_{k(N-1)} \circ \cdots \circ \tilde{T}_{k(1)} \circ \boldsymbol{\gamma}^{(0)}) - \tilde{T}_{k(n)} \circ \cdots \circ \tilde{T}_{k(N)} \circ \boldsymbol{\gamma}_{opt} \right) \\ & + (\tilde{T}_{k(n)} \circ \cdots \circ \tilde{T}_{k(N)} \circ \boldsymbol{\gamma}_{opt} - \boldsymbol{\gamma}_{opt}), \end{aligned}$$

we get

$$\begin{aligned}
& \|\tilde{T}_{k(n)} \circ \cdots \circ \tilde{T}_{k(1)} \circ \gamma^{(0)} - \gamma_{opt}\| \\
\leq & \|\tilde{T}_{k(n)} \circ \cdots \circ \tilde{T}_{k(N)} \circ (\tilde{T}_{k(N-1)} \circ \cdots \circ \tilde{T}_{k(1)} \circ \gamma^{(0)}) - \tilde{T}_{k(n)} \circ \cdots \circ \tilde{T}_{k(N)} \circ \gamma_{opt}\| \\
& + \|\tilde{T}_{k(n)} \circ \cdots \circ \tilde{T}_{k(N)} \circ \gamma_{opt} - \gamma_{opt}\| \\
\leq & (\|T_{k(n)}\| \cdots \|T_{k(N)}\|) \cdot \|\tilde{T}_{k(N-1)} \circ \cdots \circ \tilde{T}_{k(1)} \circ \gamma^{(0)} - \gamma_{opt}\| \\
& + \|\tilde{T}_{k(n)} \circ \cdots \circ \tilde{T}_{k(N)} \circ \gamma_{opt} - \gamma_{opt}\| \triangleq \text{I} \cdot \text{II} + \text{III}
\end{aligned}$$

That is,

$$\|\tilde{T}_{k(n)} \circ \cdots \circ \tilde{T}_{k(1)} \circ \gamma^{(0)} - \gamma_{opt}\| \leq \text{I} \cdot \text{II} + \text{III} \quad (3.76)$$

in short. What's more,

$$\begin{aligned}
& \tilde{T}_{k(N+M)} \circ \cdots \circ \tilde{T}_{k(N)} \circ \gamma_{opt} - \gamma_{opt} \\
= & \left( \tilde{T}_{k(N+M)} \circ \cdots \circ \tilde{T}_{k(N)} \circ \gamma_{opt} - \tilde{T}_{k(N+M)} \circ \cdots \circ \tilde{T}_{k(N)} \circ \gamma(k(N)) \right) \\
& + \left( \tilde{T}_{k(N+M)} \circ \cdots \circ \tilde{T}_{k(N+1)} \circ \gamma(k(N)) - \gamma_{opt} \right) \\
= & \left( \tilde{T}_{k(N+M)} \circ \cdots \circ \tilde{T}_{k(N)} \circ \gamma_{opt} - \tilde{T}_{k(N+M)} \circ \cdots \circ \tilde{T}_{k(N)} \circ \gamma(k(N)) \right) \\
& + \sum_{j=1}^M \left\{ \tilde{T}_{k(N+M)} \circ \cdots \circ \tilde{T}_{k(N+j)} \circ \gamma(k(N+j-1)) - \tilde{T}_{k(N+M)} \circ \cdots \circ \tilde{T}_{k(N+j)} \circ \gamma(k(N+j)) \right\} \\
& + (\gamma(k(N+M)) - \gamma_{opt}).
\end{aligned}$$

Hence

$$\text{III} \leq 2 \sup_{j \geq N} \|\gamma(k(j)) - \gamma_{opt}\| + \sum_{j \geq N} \|\gamma(k(j)) - \gamma(k(j+1))\|. \quad (3.77)$$

This decomposition is used by Wrinkler in studying simulated annealing [54].

Since  $\Sigma$  is nonsingular,

$$\text{I} \leq \prod_{j=N}^n \left( 1 - \frac{\omega \rho_0}{k^2(j)} \right) = \exp \left( \sum_N^n \log \left( 1 - \frac{\omega \rho_0}{k^2(j)} \right) \right) \leq \exp \left( - \sum_N^n \frac{1}{k^2(j)} \cdot \omega \rho_0 \right). \quad (3.78)$$

If we can show

$$\sum_1^\infty \|\gamma(k(j)) - \gamma(k(j+1))\| \text{ converges,} \quad (3.79)$$

then since  $k(j) \rightarrow \infty$ ,  $\exists N$  such that  $\sup_{j \geq N} \|\gamma(k(j)) - \gamma_{opt}\|$ ,  $\sum_N^{\infty} \|\gamma(k(j)) - \gamma(k(j+1))\|$ , and thus III, are small enough. For this  $N$ ,  $\exists M$  such that  $\sum_N^{N+M} \frac{1}{k^2(j)}$  is large enough to make sure I · II is small enough. So any cooling schedule satisfying

$$\sum_{j=1}^{\infty} \frac{1}{k^2(j)} = \infty, \text{ and } k(j) \rightarrow \infty,$$

guarantees the convergence to the optimal point  $\gamma_{opt}$ .

In the remainder, we will prove (3.79). It is enough to show

**Lemma 6**  $\|\gamma(k) - \gamma(k')\|_2 \leq \left(\frac{1}{k^2} - \frac{1}{k'^2}\right) \cdot C$  for  $\forall k, k' : k \leq k'$ .

We still consider the general relaxation form (II), with  $0 < \omega < 2$ .

$$\begin{aligned} \|\gamma(k') - \gamma(k)\|_2 &\leq \|\gamma(k') - \tilde{T}_k \circ \gamma(k') + \tilde{T}_k \circ \gamma(k') - \gamma(k)\|_2 \\ &\leq \|\gamma(k') - \tilde{T}_k \circ \gamma(k')\|_2 + \|\tilde{T}_k \circ \gamma(k') - \tilde{T}_k \circ \gamma(k)\|_2 \\ &\leq \|\tilde{T}_{k'} \circ \gamma(k') - \Theta_k \circ T_{k'} \circ \gamma(k') + \Theta_k \circ T_{k'} \circ \gamma(k') - \tilde{T}_k \circ \gamma(k')\|_2 \\ &\quad + \|T_k \circ \gamma(k') - T_k \circ \gamma(k)\|_2 \\ &\leq \|\Theta_{k'} \circ (T_{k'} \circ \gamma(k')) - \Theta_k \circ (T_{k'} \circ \gamma(k'))\|_2 \\ &\quad + \|\Theta_k \circ (T_{k'} \circ \gamma(k')) - \Theta_k \circ (T_k \circ \gamma(k'))\|_2 + \|T_k \circ \gamma(k') - T_k \circ \gamma(k)\|_2 \\ &\leq \|\Theta_{k'} \circ (T_{k'} \circ \gamma(k')) - \Theta_k \circ (T_{k'} \circ \gamma(k'))\|_2 + \|T_{k'} \circ \gamma(k') - T_k \circ \gamma(k')\|_2 \\ &\quad + \|T_k \circ \gamma(k') - T_k \circ \gamma(k)\|_2 \triangleq \text{I}^* + \text{II}^* + \text{III}^* \end{aligned}$$

That is,

$$\|\gamma(k') - \gamma(k)\|_2 \leq \text{I}^* + \text{II}^* + \text{III}^* \quad (3.80)$$

It is easy to verify

$$|\Theta_{k'} \mathbf{v} - \Theta_k \mathbf{v}| \preceq \lambda \omega \left( \frac{1}{k^2} - \frac{1}{k'^2} \right),$$

where ' $\preceq$ ' means the component-wise ' $\leq$ '. So

$$\text{I}^* \leq C \cdot \left( \frac{1}{k^2} - \frac{1}{k'^2} \right). \quad (3.81)$$

Using Fact 3), we have

$$\Pi^* = \left\| \left( \frac{1}{k^2} - \frac{1}{k'^2} \right) (\omega \cdot \mathbf{U} \mathbf{A} \mathbf{U}^T \boldsymbol{\gamma}(k') - \mathbf{H}^T \mathbf{X} \mathbf{y}) \right\|_2 \leq \left( \frac{1}{k^2} - \frac{1}{k'^2} \right) \cdot C. \quad (3.82)$$

And for  $\text{III}^*$ ,

$$\begin{aligned} \text{III}^{*2} &= \|\mathbf{J}_k (\boldsymbol{\gamma}(k') - \boldsymbol{\gamma}(k))\|_2^2 = \|(\mathbf{I} - \omega \mathbf{K}^T \mathbf{K}) (\boldsymbol{\gamma}(k') - \boldsymbol{\gamma}(k))\|_2^2 \\ &= \left\| \left( \mathbf{U} \left( \mathbf{I} - \frac{\omega}{k^2} \mathbf{A} \right) \mathbf{U}^T + (1 - \omega) \mathbf{U}_\perp \mathbf{U}_\perp^T \right) \cdot (\boldsymbol{\gamma}(k') - \boldsymbol{\gamma}(k)) \right\|_2^2 \\ &= (\boldsymbol{\gamma}(k') - \boldsymbol{\gamma}(k))^T \cdot \mathbf{U} \left( \mathbf{I} - \frac{\omega}{k^2} \mathbf{A} \right)^2 \mathbf{U}^T \cdot (\boldsymbol{\gamma}(k') - \boldsymbol{\gamma}(k)) \\ &\quad + (1 - \omega)^2 (\boldsymbol{\gamma}(k') - \boldsymbol{\gamma}(k))^T \cdot \mathbf{U}_\perp \mathbf{U}_\perp^T \cdot (\boldsymbol{\gamma}(k') - \boldsymbol{\gamma}(k)) \end{aligned}$$

Hence,

$$\text{III}^* \leq \left( \left( 1 - \omega \frac{\epsilon}{k^2} \right)^2 \|\mathbf{U} \mathbf{U}^T (\boldsymbol{\gamma}(k') - \boldsymbol{\gamma}(k))\|_2^2 + (1 - \omega)^2 \|\mathbf{U}_\perp \mathbf{U}_\perp^T (\boldsymbol{\gamma}(k') - \boldsymbol{\gamma}(k))\|_2^2 \right)^{1/2}, \quad (3.83)$$

for some  $\epsilon > 0$ , because  $\boldsymbol{\Sigma}$  and thus  $\mathbf{A}$  are nonsingular.

Summarizing (3.81), (3.82), (3.83), we obtain

$$\sqrt{\tau_1^2 + \tau_2^2} - \sqrt{\left( 1 - \frac{\omega \epsilon}{k^2} \right)^2 \tau_1^2 + (1 - \omega)^2 \tau_2^2} \leq C \cdot \left( \frac{1}{k^2} - \frac{1}{k'^2} \right)$$

where  $\tau_1 = \|\mathbf{U} \mathbf{U}^T (\boldsymbol{\gamma}(k') - \boldsymbol{\gamma}(k))\|_2$ ,  $\tau_2 = \|\mathbf{U}_\perp \mathbf{U}_\perp^T (\boldsymbol{\gamma}(k') - \boldsymbol{\gamma}(k))\|_2$ .

Using Lemma 5 and the fact that  $0 < \omega < 2$ , we get

$$\begin{aligned} \sqrt{\tau_1^2 + \tau_2^2} - \sqrt{\left( 1 - \frac{\omega \epsilon}{k^2} \right)^2 \tau_1^2 + (1 - \omega)^2 \tau_2^2} &= \frac{\tau_1^2 + \tau_2^2 - \left( 1 - \frac{\omega \epsilon}{k^2} \right)^2 \tau_1^2 - (1 - \omega)^2 \tau_2^2}{\sqrt{\tau_1^2 + \tau_2^2} + \sqrt{\left( 1 - \frac{\omega \epsilon}{k^2} \right)^2 \tau_1^2 + (1 - \omega)^2 \tau_2^2}} \\ &\geq \frac{\epsilon' \tau_2^2}{2\sqrt{\tau_1^2 + \tau_2^2}} \geq \epsilon'' \cdot \tau_2 \end{aligned}$$

for some  $\epsilon', \epsilon'' > 0$ . Hence

$$\|\mathbf{U}_\perp \mathbf{U}_\perp^T (\boldsymbol{\gamma}(k') - \boldsymbol{\gamma}(k))\| \leq C \cdot \left( \frac{1}{k^2} - \frac{1}{k'^2} \right)$$

By Lemma 5 again, Lemma 6 is true. Now the proof of Theorem 7 is complete.  $\blacksquare$

• **Proof of Proposition 5**

First (3.39) can be rewritten using the introduced operators:

$$\boldsymbol{\xi}^{(j+1)} = ((1 - \omega)I + \omega\bar{T}_k) \circ \boldsymbol{\xi}^{(j)}. \quad (3.84)$$

Obviously,  $\bar{T}_k$  is nonexpansive. We claim that the set of fixed points of  $\bar{T}_k$ , say  $F$ , is nonempty. In fact, let  $\boldsymbol{\gamma}$  be a minimizer of the convex function  $\Phi_k$  defined by (3.59). The KKT optimality condition [48] (the nonsmooth version) gives

$$\boldsymbol{\gamma} = \Theta_k \circ (\mathbf{J}_k \boldsymbol{\gamma} + \boldsymbol{\alpha}_k) = \tilde{T}_k \circ \boldsymbol{\gamma}.$$

Let  $\boldsymbol{\xi} = \mathbf{J}_k \boldsymbol{\gamma} + \boldsymbol{\alpha}_k$ . Then  $\boldsymbol{\xi} = \mathbf{J}_k(\Theta_k \circ \boldsymbol{\xi}) + \boldsymbol{\alpha}_k = \bar{T}_k \circ \boldsymbol{\xi}$ . So  $\bar{T}_k$  has at least one fixed point. In the rest of this proof, all subscripts  $k$  are abbreviated for simplicity.

For  $0 < \omega < 1$ , (3.84) is the Mann iterates [35] introduced for nonexpansive mapping  $\bar{T}$ . And it is known to converge to a fixed point of  $\bar{T}$  if  $F$  is nonempty; see Opial [41], Browder and Petryshyn [12], or Dotson [21].

Now consider  $1 < \omega < 2$ . Let  $\omega = 1 + \omega'$ . So  $\omega' \in (0, 1)$  and

$$\boldsymbol{\xi}^{(j+1)} = \omega'(2\bar{T} - I) \circ \boldsymbol{\xi}^{(j)} + (1 - \omega')\bar{T} \circ \boldsymbol{\xi}^{(j)}$$

If  $2\bar{T} - I$  is nonexpansive,  $(1 - \omega)I + \omega\bar{T}$  is nonexpansive for any  $\omega \in (1, 2)$ .

Let  $\boldsymbol{\xi} \in F$ . Clearly,  $\bar{T} \circ \boldsymbol{\xi} = \boldsymbol{\xi} = (2\bar{T} - I) \circ \boldsymbol{\xi}$ . On the one hand,

$$\begin{aligned} \|\boldsymbol{\xi}^{(j+1)} - \boldsymbol{\xi}\|_2^2 &= \left\| \omega' \left( (2\bar{T} - I) \circ \boldsymbol{\xi}^{(j)} - \boldsymbol{\xi} \right) + (1 - \omega') \left( \bar{T} \circ \boldsymbol{\xi}^{(j)} - \boldsymbol{\xi} \right) \right\|_2^2 \\ &\leq \omega'^2 \|\boldsymbol{\xi}^{(j)} - \boldsymbol{\xi}\|_2^2 + (1 - \omega')^2 \|\boldsymbol{\xi}^{(j)} - \boldsymbol{\xi}\|_2^2 \\ &\quad + 2\omega'(1 - \omega') \langle (2\bar{T} - I) \circ \boldsymbol{\xi}^{(j)} - \boldsymbol{\xi}, \bar{T} \circ \boldsymbol{\xi}^{(j)} - \boldsymbol{\xi} \rangle. \end{aligned}$$

On the other hand,

$$\begin{aligned} a^2 \|\boldsymbol{\xi}^{(j)} - \bar{T} \circ \boldsymbol{\xi}^{(j)}\|_2^2 &= a^2 \|(2\bar{T} - I) \circ \boldsymbol{\xi}^{(j)} - \bar{T} \circ \boldsymbol{\xi}^{(j)}\|_2^2 \\ &= a^2 \left\| \left( (2\bar{T} - I) \circ \boldsymbol{\xi}^{(j)} - \boldsymbol{\xi} \right) - \left( \bar{T} \circ \boldsymbol{\xi}^{(j)} - \boldsymbol{\xi} \right) \right\|_2^2 \\ &\leq a^2 \|\boldsymbol{\xi}^{(j)} - \boldsymbol{\xi}\|_2^2 + a^2 \|\boldsymbol{\xi}^{(j)} - \boldsymbol{\xi}\|_2^2 \\ &\quad - 2a^2 \langle (2\bar{T} - I) \circ \boldsymbol{\xi}^{(j)} - \boldsymbol{\xi}, \bar{T} \circ \boldsymbol{\xi}^{(j)} - \boldsymbol{\xi} \rangle. \end{aligned}$$

Letting  $a^2 = \omega'(1 - \omega')$ , we obtain

$$\|\boldsymbol{\xi}^{(j+1)} - \boldsymbol{\xi}\|_2^2 + \omega'(1 - \omega')\|\boldsymbol{\xi}^{(j)} - \bar{T} \circ \boldsymbol{\xi}^{(j)}\|_2^2 \leq \|\boldsymbol{\xi}^{(j)} - \boldsymbol{\xi}\|_2^2,$$

and so

$$\|\boldsymbol{\xi}^{(j+1)} - \boldsymbol{\xi}^{(j)}\|_2^2 = \omega^2\|\boldsymbol{\xi}^{(j)} - \bar{T} \circ \boldsymbol{\xi}^{(j)}\|_2^2 \leq \frac{\omega^2}{(\omega - 1)(2 - \omega)} \cdot \left( \|\boldsymbol{\xi}^{(j)} - \boldsymbol{\xi}\|_2^2 - \|\boldsymbol{\xi}^{(j+1)} - \boldsymbol{\xi}\|_2^2 \right).$$

It follows that  $\sum \|\boldsymbol{\xi}^{(j+1)} - \boldsymbol{\xi}^{(j)}\|_2^2$  converges. Note that we only used quasi-nonexpansiveness [21] in the above proof.

Hence  $(1 - \omega)I + \omega\bar{T}$  is asymptotically regular – in fact, it is a reasonable wanderer [12]. And  $\boldsymbol{\xi}^{(j)}$ , or  $\boldsymbol{\gamma}^{(j)}$ , converges by Opial's classical work [41]. ■

• **Proof of Proposition 6**

The SVD for  $\mathbf{F}_1$  is well known (see, for example, [2] for a detailed derivation).

Consider a  $d$ -by- $d$  matrix  $\mathbf{E}$  of all ones:  $\mathbf{E} = \mathbf{1} \cdot \mathbf{1}^T$ . It is easy to diagonalize  $\mathbf{E}$ . First,

$$\mathbf{E} \cdot \begin{bmatrix} \mathbf{1} & \mathbf{F}_1^T \end{bmatrix} = \begin{bmatrix} \mathbf{1} & \mathbf{F}_1^T \end{bmatrix} \cdot \text{diag}\{d, 0, \dots, 0\}.$$

So  $\mathbf{F}_1^T \perp \mathbf{1}$ , i.e.,  $\mathbf{F}_1 \mathbf{1} = \mathbf{0}$ . It follows that  $\mathbf{V}_1^T \mathbf{1} = \mathbf{D}_1^{-1} \mathbf{U}_1^T \mathbf{F}_1 \mathbf{1} = \mathbf{0}$ , and  $\tilde{\mathbf{V}}_1$  is orthonormal. Hence  $\mathbf{E} = \tilde{\mathbf{V}}_1 \text{diag}\{d, 0, \dots, 0\} \mathbf{V}_1^T$ .

For  $\mathbf{T}_1 = \begin{bmatrix} \mathbf{I} \\ \lambda \mathbf{F}_1 \end{bmatrix}$ , we have

$$\mathbf{T}_1^T \mathbf{T}_1 = \mathbf{I} + \lambda^2 \mathbf{V}_1 \mathbf{D}_1^2 \mathbf{V}_1^T = \tilde{\mathbf{V}}_1^T \tilde{\mathbf{V}}_1 + \lambda^2 \mathbf{V}_1 \mathbf{D}_1^2 \mathbf{V}_1^T = \tilde{\mathbf{V}}_1 \begin{bmatrix} 1 & \\ & \mathbf{I} + \lambda^2 \mathbf{D}_1^2 \end{bmatrix} \tilde{\mathbf{V}}_1^T.$$

On the other hand,

$$\begin{aligned} \mathbf{T}_1 \tilde{\mathbf{V}}_1 \begin{bmatrix} 1 & \\ & \mathbf{I} + \lambda^2 \mathbf{D}_1^2 \end{bmatrix}^{-\frac{1}{2}} &= \begin{bmatrix} \mathbf{I} \\ \lambda \mathbf{F}_1 \end{bmatrix} \cdot \begin{bmatrix} \frac{1}{\sqrt{d}} \mathbf{1} & \mathbf{V}_1 \end{bmatrix} \cdot \begin{bmatrix} 1 & \\ & (\mathbf{I} + \lambda^2 \mathbf{D}_1^2)^{-\frac{1}{2}} \end{bmatrix} \\ &= \begin{bmatrix} \frac{1}{\sqrt{d}} \mathbf{1} & \mathbf{V}_1 \\ \frac{\lambda}{\sqrt{d}} \cdot \mathbf{0} & \lambda \mathbf{U}_1 \mathbf{D}_1 \end{bmatrix} \cdot \begin{bmatrix} 1 & \\ & (\mathbf{I} + \lambda^2 \mathbf{D}_1^2)^{-\frac{1}{2}} \end{bmatrix} = \tilde{\mathbf{U}}_1. \end{aligned}$$

For  $\mathbf{F}_2$ ,  $\mathbf{F}_2^T \mathbf{F}_2 = d\mathbf{I} - \mathbf{1} \cdot \mathbf{1}^T = \tilde{\mathbf{V}}_1 \text{diag}\{0, d, \dots, d\} \tilde{\mathbf{V}}_1^T$ . So  $\mathbf{D}_2 = \text{diag}\{0, \sqrt{d}, \dots, \sqrt{d}\}$ , and if we take  $\mathbf{V}_2 = \tilde{\mathbf{V}}_1$ ,  $\mathbf{U}_2 = \begin{bmatrix} \mathbf{u}_{21} & \dots & \mathbf{u}_{2d} \end{bmatrix}$  satisfies  $\mathbf{F}_2 \tilde{\mathbf{V}}_1 = \mathbf{U}_2 \mathbf{D}_2$ . It implies  $\begin{bmatrix} \mathbf{u}_{22} & \dots & \mathbf{u}_{2d} \end{bmatrix} = \frac{1}{\sqrt{d}} \mathbf{F}_2 \mathbf{V}_1$ .  $\mathbf{u}_{21}$  is a normalized eigenvector of  $\mathbf{F}_2 \mathbf{F}_2^T$  corresponding to eigenvalue 0. It is easy to verify that

$$\begin{bmatrix} 0 & \dots & 0 & 1 & -1 & 1 \end{bmatrix}^T / \sqrt{3}$$

is one choice.

Finally, for  $\mathbf{T}_2$ ,  $\mathbf{T}_2^T \mathbf{T}_2 = \mathbf{I} + \lambda^2 \mathbf{F}_2^T \mathbf{F}_2 = \mathbf{I} + \lambda^2 \mathbf{V}_2 \mathbf{D}_2^2 \mathbf{V}_2^T = \mathbf{V}_2 (\mathbf{I} + \lambda^2 \mathbf{D}_2^2) \mathbf{V}_2^T$ . Moreover,  $\mathbf{T}_2 \mathbf{V}_2 (\mathbf{I} + \lambda^2 \mathbf{D}_2^2)^{-\frac{1}{2}} = \begin{bmatrix} \mathbf{I} \\ \lambda \mathbf{F}_2 \end{bmatrix} \mathbf{V}_2 (\mathbf{I} + \lambda^2 \mathbf{D}_2^2)^{-\frac{1}{2}} = \begin{bmatrix} \mathbf{V}_2 \\ \lambda \mathbf{U}_2 \mathbf{D}_2 \end{bmatrix} (\mathbf{I} + \lambda^2 \mathbf{D}_2^2)^{-\frac{1}{2}} = \tilde{\mathbf{U}}_2$ . ■

## Chapter 4

# Thresholding-based Iterative Selection Procedures for Model Selection and Shrinkage

### 4.1 Motivation

#### 4.1.1 From orthogonal designs to non-orthogonal designs

We consider the penalized regression problem

$$\min_{\boldsymbol{\beta}} \frac{1}{2} \|\mathbf{X}\boldsymbol{\beta} - \mathbf{y}\|_2^2 + P(\boldsymbol{\beta}; \lambda) (\triangleq f(\boldsymbol{\beta})), \quad (4.1)$$

where  $\mathbf{X}$  is the regression matrix,  $\mathbf{y}$  is the response vector, and  $P(\boldsymbol{\beta}; \lambda)$  represents the penalty with  $\lambda$  as the regularization parameter. Here  $p$  may be greater than  $n$ . In this chapter, we assume  $\boldsymbol{\beta}$  is sparse, and use (4.1) to solve the variable selection problem.

If  $P(\boldsymbol{\beta}; \lambda) = \lambda \|\boldsymbol{\beta}\|_1$ , then (4.1) is the lasso [50], a basic and popular method in variable selection. However, although the  $l_1$ -norm provides the best convex approximation to the  $l_0$ -norm, the lasso results in inconsistent selection (cf. the irrepresentable conditions [60]) and introduces extra bias in estimation [38].

On the other hand, if we concentrate on orthogonal designs only, i.e.,  $\mathbf{X}^T \mathbf{X} = \mathbf{I}$ , like in wavelets, there are rich theories and algorithms for various types of penalties, such as the SCAD [5], the transformed  $L_1$  [31], and the  $L_0$ -penalty [25]. In particular, (a) the fitting

part of the penalized regression (4.1) is **separable** in this case, which means we only need to deal with the univariate case, if  $P$  is also separable (which is true in general); (b) even if  $P$  is **nonconvex**, it still often results in a *unique* solution.

One of our main goals in this chapter is to borrow these rich results in the orthogonal design to help us solve the general problem (4.1). We use the following mechanism to achieve this. Define

$$g(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \frac{1}{2} \|\mathbf{X}\boldsymbol{\gamma} - \mathbf{y}\|_2^2 + P(\boldsymbol{\gamma}; \lambda) + \frac{1}{2} \langle (\mathbf{I} - \boldsymbol{\Sigma})\boldsymbol{\beta}, \boldsymbol{\beta} - 2\boldsymbol{\gamma} \rangle \quad (4.2)$$

Here  $\langle \mathbf{a}, \mathbf{b} \rangle = \mathbf{a}^T \mathbf{b}$ ,  $\boldsymbol{\Sigma} = \mathbf{X}^T \mathbf{X}$ .

Given  $\boldsymbol{\beta}$ , minimizing  $g$  over  $\boldsymbol{\gamma}$  is equivalent to

$$\min_{\boldsymbol{\gamma}} \frac{1}{2} \|\boldsymbol{\gamma} - [(\mathbf{I} - \boldsymbol{\Sigma})\boldsymbol{\beta} + \mathbf{X}^T \mathbf{y}]\|_2^2 + P(\boldsymbol{\gamma}; \lambda). \quad (4.3)$$

In contrast to (4.1), this problem has an orthogonal design — as mentioned earlier this is easier to handle both in computation and in theory. For example, we may adopt some nonconvex penalties, and they still result in a unique solution of  $\boldsymbol{\gamma}$ .

Given  $\boldsymbol{\gamma}$ , minimizing  $g$  over  $\boldsymbol{\beta}$  is equivalent to

$$\min_{\boldsymbol{\beta}} \frac{1}{2} \langle (\mathbf{I} - \boldsymbol{\Sigma})\boldsymbol{\beta}, \boldsymbol{\beta} - 2\boldsymbol{\gamma} \rangle. \quad (4.4)$$

Taking its derivative with respect to  $\boldsymbol{\beta}$  gives  $(\mathbf{I} - \boldsymbol{\Sigma})(\boldsymbol{\beta} - \boldsymbol{\gamma}) = \mathbf{0}$ , from which it follows that  $\boldsymbol{\beta} = \boldsymbol{\gamma}$  if  $\|\boldsymbol{\Sigma}\|_2 < 1$ . Note that (4.4) is a convex optimization. Therefore, the optimal value of  $g$  is always achieved at  $\boldsymbol{\gamma} = \boldsymbol{\beta}$ .

The connection to the original problem is now clear: it is easy to verify  $\min_{\boldsymbol{\beta}} g(\boldsymbol{\beta}, \boldsymbol{\beta})$  is equivalent to  $\min_{\boldsymbol{\beta}} f(\boldsymbol{\beta})$ . The advantage of optimizing  $g$  instead of  $f$  is that given  $\boldsymbol{\beta}$ , the problem is orthogonal and separable in  $\boldsymbol{\gamma}$ , and we can adopt far more flexible penalties in the algorithm design, including the nonconvex ones.

## 4.2 Thresholding-based Iterative Selection Procedures (TISP)

### 4.2.1 Thresholding rules and penalties

As the title suggests, our starting point in this chapter is thresholding rules rather than different forms of the penalty function. One direct reason is that different  $P$ 's may result in

the same estimator and the same thresholding, say, in the situation of hard-thresholding [3, 25]. Moreover, starting with thresholding functions facilitates the computation (as will be shown in the next subsection). Besides, there is also a universal connection between thresholding rules and penalty functions that we will investigate in this subsection. For convenience, we consider the univariate case only.

A thresholding function, denoted by  $\Theta(\cdot; \lambda)$ , with  $\lambda$  as a parameter, is required to satisfy:

1.  $\Theta(\cdot; \lambda)$  is an odd function. ( $\Theta_+(\cdot; \lambda)$  is used to denote the  $\Theta(\cdot; \lambda)$  restricted to  $R_+ = [0, \infty)$ .)
2.  $\Theta$  is a shrinkage rule:  $0 \leq \Theta_+(t; \lambda) \leq t, \forall t \in R_+$ .
3.  $\Theta_+$  is nondecreasing on  $R_+$ , and  $\Theta_+(t; \lambda) \rightarrow \infty$  as  $t \rightarrow \infty$ .

In addition, it is natural to have  $\Theta_+(t; \lambda) = 0, 0 \leq t \leq \tau$  for some  $\tau \geq 0$ .

Given a thresholding rule  $\Theta(\cdot; \lambda)$ , define

$$\Theta^{-1}(u; \lambda) = \sup\{t : \Theta(t; \lambda) \leq u\} \text{ and } \Theta^{-1}(-u; \lambda) = -\Theta^{-1}(u; \lambda),$$

for any  $u \in R_+$ . And

$$s(u; \lambda) \triangleq \Theta^{-1}(u; \lambda) - u, \forall u. \tag{4.5}$$

Let  $P$  be a continuous and positive penalty defined by

$$P(\theta; \lambda) = \int_0^{|\theta|} s(u; \lambda) du. \tag{4.6}$$

Antoniadis [4] showed the following result for this constructed  $P$ .

**Proposition 7** *The minimization problem  $\min_{\theta} (t - \theta)^2/2 + P(\theta; \lambda)$  has a unique optimal solution  $\hat{\theta} = \Theta(t; \lambda)$  for every  $t$  at which  $\Theta(\cdot; \lambda)$  is continuous.*

By the way, if we define

$$\psi(t) = t - \Theta(t), \tag{4.7}$$

then it is Huber's  $\psi$ -function in M-estimates; see [29, 4].

Note that (4.6) is not the only way to construct a penalty that leads to  $\Theta$  in solving the optimization. For example, in the situation of hard-thresholding, in addition to the

continuous  $P = \lambda^2/2 - (|\theta| - \lambda)^2 1_{|\theta| < \lambda}/2$  constructed via (4.6),

$$P(\theta; \lambda) = \begin{cases} \lambda|\theta|, & \text{if } |\theta| < \lambda \\ \lambda^2/2, & \text{if } |\theta| \geq \lambda \end{cases}, \text{ and } P(\theta; \lambda) = \frac{\lambda^2}{2} \cdot 1_{\theta \neq 0} \quad (4.8)$$

are also valid choices [3, 25].

### 4.2.2 TISP and its convergence

Now we go back to the mechanism introduced in Section 4.1 for the penalized regression problem (4.1), with  $P$  constructed from a given thresholding function  $\Theta$ . Solving (4.3) yields  $\boldsymbol{\gamma} = \Theta((\mathbf{I} - \boldsymbol{\Sigma})\boldsymbol{\beta} + \mathbf{X}^T \mathbf{y}; \lambda)$ . Seen from (4.4), our iterates simplify to

$$\boldsymbol{\beta}^{(j+1)} = \Theta((\mathbf{I} - \boldsymbol{\Sigma})\boldsymbol{\beta}^{(j)} + \mathbf{X}^T \mathbf{y}; \lambda). \quad (4.9)$$

This iterative procedure is referred to as the **Thresholding-based Iterative Selection Procedure (TISP)**. As discussed earlier, it provides a feasible way to tackle the original optimization (4.1).

There are rich examples for the procedure defined by (4.9). Using a soft-thresholding in (4.9), we immediately obtain the iterative algorithm [18] (in vector form) for solving the lasso with  $P(\boldsymbol{\beta}; \lambda) = \lambda \|\boldsymbol{\beta}\|_1$ . The corresponding pathwise algorithm has been considered to be the fastest in solving the lasso problem to date, especially when  $p > n$  [27]. If we substitute hard-thresholding for  $\Theta$ , seen from (4.8), it is an alternative optimization for the penalized regression with

$$P = c \cdot \sum_i 1_{\beta_i \neq 0} = c \cdot \|\boldsymbol{\beta}\|_0,$$

i.e., the  $l_0$ -penalized regression problem. We can also replace the hard-thresholding by the more smoothed SCAD to reduce instability. Finally, it is worth mentioning that TISP may also include the ridge penalty  $P(\boldsymbol{\beta}; \lambda) = \lambda \|\boldsymbol{\beta}\|_2^2/2$ , if we set

$$\Theta(t; \lambda) = \frac{t}{1 + \lambda}, \quad (4.10)$$

thanks to the generic definition of a thresholding function.

Obviously, if  $\boldsymbol{\Sigma}$  is nonsingular, and so  $n > p$ , the TISP mapping is a contraction and

thus the sequence  $\boldsymbol{\beta}^{(j)}$  converges to a stationary point of (4.1). In contrast, a great difficulty encountered when  $\boldsymbol{\Sigma}$  is singular is that TISP may not be a *nonexpansive* operator<sup>1</sup> for most thresholdings (except soft-thresholding), let alone a contraction. The following studies cover the large  $p$  case ( $p > n$ ). We use  $\mu(\mathbf{A})$  to represent an arbitrary singular value of matrix  $\mathbf{A}$ , and  $\mu_{\max}(\mathbf{A})$  ( $\mu_{\min}(\mathbf{A})$ ) the max (min) of  $\mu(\mathbf{A})$ , respectively.

**Definition 3 (Bounded curvature condition)** *The penalty function defined by (4.6) fulfills the bounded curvature condition (BCC) if for some positive semi-definite matrix  $\mathbf{H}$ ,*

$$P(\boldsymbol{\beta} + \boldsymbol{\Delta}; \lambda) \geq P(\boldsymbol{\beta}; \lambda) + \langle \boldsymbol{\Delta}, \mathbf{s} \rangle - \frac{1}{2} \boldsymbol{\Delta}^T \mathbf{H} \boldsymbol{\Delta}. \quad (4.11)$$

where  $\mathbf{s} = \mathbf{s}(\boldsymbol{\beta}; \lambda)$  is given by (4.5).

**Theorem 8** *Given the TISP (4.9), suppose the penalty function defined by (4.6) satisfies the bounded curvature condition (BCC) for some positive semi-definite  $\mathbf{H}$ . Then if  $\mu_{\max}(\boldsymbol{\Sigma}) \leq 1 \vee (2 - \mu_{\max}(\mathbf{H}))$ ,*

$$f(\boldsymbol{\beta}^{(j)}) \geq f(\boldsymbol{\beta}^{(j+1)}). \quad (4.12)$$

Moreover, if  $\mu_{\max}(\boldsymbol{\Sigma}) < 1 \vee (2 - \mu_{\max}(\mathbf{H}))$ , there exists a constant  $C > 0$ , dependent on  $\mathbf{X}$ ,  $\mathbf{H}$  only, such that

$$f(\boldsymbol{\beta}^{(j)}) - f(\boldsymbol{\beta}^{(j+1)}) \geq C \cdot \|\boldsymbol{\beta}^{(j)} - \boldsymbol{\beta}^{(j+1)}\|_2^2. \quad (4.13)$$

Therefore, for an arbitrary  $\mathbf{X}$ , we can use TISP of the following form in practice

$$\boldsymbol{\beta}^{(j+1)} = \Theta \left( \left( \mathbf{I} - \frac{1}{k_0^2} \boldsymbol{\Sigma} \right) \boldsymbol{\beta}^{(j)} + \frac{1}{k_0^2} \mathbf{X}^T \mathbf{y}; \frac{\lambda}{k_0^2} \right), \quad (4.14)$$

where  $k_0 = \mu_{\max}(\mathbf{X}) = \|\mathbf{X}\|_2$ . The BCC is not a restrictive condition. For example, for soft-thresholding,  $\mathbf{H} = \mathbf{0}$  since  $\|\boldsymbol{\beta}\|_1$  is convex; for hard-thresholding,  $\mathbf{H} = \mathbf{I}$ ; for SCAD-thresholding, we can take  $\mathbf{H} = \mathbf{I}/(a - 1)$  — note that the parameter  $a$  is assumed to be greater than 2, and so  $\mathbf{H}$  is positive definite. See Section 4.6 for details. Correspondingly, we obtain the following corollaries.

<sup>1</sup>An operator  $T$  is called nonexpansive if  $\|T(x) - T(y)\| \leq \|x - y\|$  for any  $x, y$ .

**Corollary 1** *Suppose  $\Theta$  is soft-thresholding. If  $\mu_{\max}(\mathbf{X}) < \sqrt{2}$ , then (4.13) holds.*

**Corollary 2** *Suppose  $\Theta$  is hard-thresholding. If  $\mu_{\max}(\mathbf{X}) \leq 1$ , then (4.12) holds; further, if  $\mu_{\max}(\mathbf{X}) < 1$ , then (4.13) is true.*

**Corollary 3** *Suppose  $\Theta$  is SCAD-thresholding. If  $\mu_{\max}(\mathbf{X}) < \sqrt{2 - \frac{1}{a-1}}$ , then (4.13) holds.*

Corollary 1 generalizes the lasso result by Daubechies *et al.* [18], and coincides with our previous study (see Chapter 3 or [46]). Corollary 3 covers the orthogonal case, since SCAD assumes  $a > 2$  and thus  $\sqrt{2 - \frac{1}{a-1}} > 1$ . Finally, it is worth pointing out that TISP may not always be an MM algorithm [33] like the LLA method by Zou and Li [63]. Yet Theorem 8 states that if  $\mathbf{X}$  is scaled down properly (which does not affect the variable selection),  $f(\boldsymbol{\beta}^{(j)})$  is nondecreasing all the time during the procedure.

We can easily show a result similar to Zou and Li [63]:

**Proposition 8** *Suppose  $\mu_{\max}(\boldsymbol{\Sigma}) < 1 \vee (2 - \mu_{\max}(\mathbf{H}))$ . Give an initial point  $\boldsymbol{\beta}(0)$ , if  $\boldsymbol{\beta}^*$  is a limit point of the TISP sequence  $\boldsymbol{\beta}^{(j)}$ , then  $\boldsymbol{\beta}^*$  is a stationary point of  $f(\boldsymbol{\beta})$  (4.1), or equivalently, a fixed point of (4.9).*

Denote by  $F$  the set of the fixed points of TISP. That is, given any  $\boldsymbol{\beta}^* \in F$ , it satisfies the equation

$$\boldsymbol{\beta} = \Theta((\mathbf{I} - \boldsymbol{\Sigma})\boldsymbol{\beta} + \mathbf{X}^T \mathbf{y}; \lambda). \quad (4.15)$$

Clearly, local minima of  $f$  are fixed points. In the next section, we will perform a nonasymptotic study of the good properties of the points in  $F$ . Here, we give the following optimality result.

**Proposition 9** *Let  $\boldsymbol{\beta}^* \in F$  and suppose  $\mu_{\max}(\mathbf{H}) \leq 1$ . If  $\mu_{\max}(\mathbf{H}) \leq \mu(\boldsymbol{\Sigma}) \leq 2 - \mu_{\max}(\mathbf{H})$ , then  $\boldsymbol{\beta}^*$  is a global minimizer of  $f$ .*

Although the fact that nonconvex penalties often result in a unique optimal solution in the *orthogonal* design is well known, this proposition states novelly that the same conclusion holds as long as  $\mathbf{X}$  is not too far from orthogonal (characterized in terms of  $\mathbf{H}$ ). For instance, for SCAD thresholding and penalty, TISP must lead to the global minimum of  $f$ , provided  $\frac{1}{\sqrt{a-1}} \leq \mu(\mathbf{X}) \leq \sqrt{2 - \frac{1}{a-1}}$ , or  $0.61 \leq \mu(\mathbf{X}) \leq 1.27$  when  $a = 3.7$  (the default choice in

SCAD), given any initial point  $\beta^{(0)}$ . In summary, TISP is a successful algorithm for solving the penalized regressions for a general design matrix.

### 4.3 Selection and Estimation via TISP

TISP provides a very simple way to do variable selection via penalized regressions. In this section, we will perform a detailed theoretical study of the variable selection and the coefficient estimation of TISPs based on different thresholdings. All our studies are nonasymptotic.

Given  $\Theta(\cdot; \lambda)$ , denote its thresholding value by  $\tau(\lambda)$ , i.e.,  $\Theta(t; \lambda) = 0 \forall t : |t| < \tau$  and  $\Theta(t; \lambda) \neq 0$  for  $|t| > \tau$ . Assume  $\tau > 0$ . To facilitate our study of TISP based on the KKT equation (4.15), we define another version of (4.5), called the generalized sign. Introduce

$$\widetilde{\text{Sgn}}(u; \lambda) = \{s \in R : \Theta(u + \tau s; \lambda) = u\} \quad \text{if } u \in \text{ran}(\Theta),$$

and  $\widetilde{\text{Sgn}}(u; \lambda) = \{0\}$  otherwise;  $\widetilde{\text{sgn}}(u; \lambda)$  is used to denote a specific element in  $\widetilde{\text{Sgn}}(u; \lambda)$ . The vector versions of  $\widetilde{\text{Sgn}}$  and  $\widetilde{\text{sgn}}$  can be defined correspondingly.

As a demonstration, if  $\Theta(\cdot; \lambda)$  is soft-thresholding,  $\tau = \lambda$  and  $\widetilde{\text{Sgn}}(\beta) = \{s : s_i = 1 \text{ if } \beta_i > 0, s_i = -1 \text{ if } \beta_i < 0, \text{ and } s_i \in [-1, 1] \text{ if } \beta_i = 0\}$ . Thus now  $\widetilde{\text{Sgn}}(\beta)$  is the subdifferential of  $\|\beta\|_1$ , and  $\widetilde{\text{sgn}}(\beta)$  is a subgradient [48]. And for hard-thresholding,  $\widetilde{\text{Sgn}}(\beta) = \{s : s_i = 0 \text{ if } \beta_i \neq 0, s_i \in [-1, 1] \text{ if } \beta_i = 0\}$ . In general we have

**Proposition 10** *Suppose  $\Theta(\cdot; \lambda)$  is sandwiched by soft- and hard-thresholdings,  $\Theta_S(\cdot; \tau)$  and  $\Theta_H(\cdot; \tau)$ , i.e.,*

$$(\Theta_S)_+(t; \tau) \leq \Theta_+(t; \lambda) \leq (\Theta_H)_+(t; \tau), \forall t \in R_+. \quad (4.16)$$

*Then  $0 \leq \widetilde{\text{sgn}}(u) \leq 1$  if  $u > 0$ ,  $-1 \leq \widetilde{\text{sgn}}(u) \leq 0$  if  $u < 0$ , and  $\widetilde{\text{sgn}}(0) \in [-1, 1]$ .*

This proposition is easy to prove from the non-decreasing property of  $\Theta$ . Throughout the rest of the section, we assume  $\Theta$  always satisfies the sandwiching condition (4.16). By the definition of the generalized signs, (4.15) is equivalent to

$$\Sigma\beta = \mathbf{X}^T \mathbf{y} - \tau \widetilde{\text{sgn}}(\beta; \lambda),$$

for some  $\widetilde{\text{sgn}}(\boldsymbol{\beta}; \lambda) \in \widehat{\text{Sgn}}(\boldsymbol{\beta}; \lambda)$ . In the following, we study the TISP estimate based on the scaled form (4.14). Let  $\hat{\boldsymbol{\beta}}$  be a fixed point of (4.14) and suppose  $\tau(\lambda) = k^2\tau(\lambda/k^2)$  for any  $k \in R$ . Then the KKT equation for this TISP estimate is

$$\boldsymbol{\Sigma}\hat{\boldsymbol{\beta}} = \mathbf{X}^T \mathbf{y} - \tau \widetilde{\text{sgn}}(\hat{\boldsymbol{\beta}}; \lambda/k_0^2). \quad (4.17)$$

Recall that  $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ ,  $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I})$ , and  $\boldsymbol{\beta}$  is sparse. Let  $z = \{i : \beta_i = 0\}$ ,  $nz = \{i : \beta_i \neq 0\}$ ,  $d_z = |z|$ ,  $d_{nz} = |nz|$ . To study the sign-consistency of a TISP estimate, we denote by  $p_s$  the probability of successful sign recovery, that is, the probability that there exists a  $\hat{\boldsymbol{\beta}} \in F$  such that  $\text{sgn}(\hat{\boldsymbol{\beta}}) = \text{sgn}(\boldsymbol{\beta})$ .

To facilitate asymptotic discussions, we assume  $\mathbf{X}$  has been scaled to have all column  $l_2$ -norms equal to  $\sqrt{n}$ . Define  $\boldsymbol{\Sigma}^{(s)} = \boldsymbol{\Sigma}/n$ . To get a better form of the bounds for  $p_s$ , we define two quantities  $\mu = \mu_{\min}(\boldsymbol{\Sigma}_{nz, nz}^{(s)})$  and  $\kappa \triangleq \max_{i \in z} \|\boldsymbol{\Sigma}_{i, nz}^{(s)}\|_2 / \sqrt{d_{nz}}$ . The following nonasymptotic result is always true regarding the selection via TISP.

**Theorem 9** *Assume  $\mu \geq \kappa d_{nz}$ ,  $\mu > 0$  and  $\min |\boldsymbol{\beta}_{nz}| \geq \frac{d_{nz}\tau}{n\mu}$ , then*

$$p_s \geq [1 - 2\Phi(-M)]^{d_z} [1 - 2\Phi(-L)]^{d_{nz}}, \quad (4.18)$$

where  $M = \left(1 - \frac{\kappa \cdot d_{nz}}{\mu}\right) \frac{\tau}{\sqrt{n}\sigma}$ ,  $L = \frac{\sqrt{\mu n}}{\sigma} \left(\min |\boldsymbol{\beta}_{nz}| - \frac{\tau d_{nz}}{\mu n}\right)$ , and  $\Phi$  is the standard normal distribution.

**Corollary 4** *Under the conditions of Theorem 9, we have*

$$1 - p_s \leq 2d_z \varphi(M)/M + 2d_{nz} \varphi(L)/L, \quad (4.19)$$

where  $\varphi$  is the standard normal density.

We can also use this theorem to explore the asymptotics. Assume  $\boldsymbol{\beta}$ ,  $d_z$ , and  $d_{nz}$  are fixed,  $n \rightarrow \infty$ , then under some regularity conditions we get: if  $\lambda(n)/\sqrt{n} \rightarrow \infty$  and  $\lambda(n)/n \rightarrow 0$ , then the TISP estimate is sign consistent. This result, in the Soft-TISP (lasso) case, coincides with other studies like [34, 60]. (Moreover, we know that under this condition,  $R_z \rightarrow 0$ ,  $R_{nz} \rightarrow 0$  by Theorem 11.)

Unfortunately, the regularity condition  $\mu \geq \kappa d_{nz}$  cannot be removed in general. In the lasso case, it is a version of the irrepresentable conditions [60]. (We took this more restrictive

form because it leads to more nice-looking bounds in (4.18) and (4.19).) However, for hard-thresholding-like  $\Theta$ 's, this is unnecessary and we can obtain stronger results.

We say that  $\Theta$  belongs to the *hard-thresholding family* if and only if

$$\Theta(t; \lambda) = t, \forall t : |t| > c \cdot \tau, \quad (4.20)$$

for some constant  $c \geq 1$ . Hard-thresholding and SCAD-thresholding are two examples with  $c = 1$ ,  $a$  respectively.

**Theorem 10** *Suppose  $\Theta$  belongs to the hard-thresholding family and  $\min |\beta_{nz}| \geq c\tau/k_0^2$ . Then*

$$p_s \geq [1 - 2\Phi(-M')]^{d_z} [1 - 2\Phi(-L')]^{d_{nz}}, \quad (4.21)$$

where  $M' = \frac{c\tau}{\sqrt{n}\sigma}$ ,  $L' = \frac{\sqrt{\mu n}}{\sigma} \left( \min |\beta_{nz}| - \frac{c\tau}{k_0^2} \right)$ .

This result is strictly better than the bound in (4.18) if  $c < d_{nz}k_0^2/(\mu n)$ , which is true in general for both hard- and scad-thresholding.

**Corollary 5** *Under the conditions of Theorem 10, we have*

$$1 - p_s \leq 2d_z\varphi(M')/M' + 2d_{nz}\varphi(L')/L'. \quad (4.22)$$

So the TISP induced by a  $\Theta$  in the hard-thresholding family can achieve better performance in variable selection.<sup>2</sup> This will be verified empirically in the next section.

We obtain the following bounds for the estimation risks of a general TISP.

**Theorem 11** *Let  $\nu = \mu_{\min}(\Sigma_{z,z}^{(s)})$  and  $\hat{\beta} \in F$ . Define  $R_{nz} = E(\|\beta_{nz} - \hat{\beta}_{nz}\|_2^2)$ , and  $R_z = E(\|\hat{\beta}_z\|_2^2)$ . Suppose  $\Sigma$  is nonsingular. Then*

$$R_{nz} \leq \frac{3}{n} \left[ \frac{d_{nz}}{\mu} \sigma^2 + \frac{d_{nz}}{\mu^2} \frac{\tau^2}{n} + \kappa^2 \frac{d_z d_{nz}}{\mu^2} \cdot n R_z \right]. \quad (4.23)$$

---

<sup>2</sup>Note that, however, the regularization parameters are generally tuned to reduce the test error.

And

$$R_z \leq \frac{\sigma^2 d_z^2}{n \nu^2} (K_1 M + K_2 \frac{1}{M}) \varphi(M), \quad (4.24)$$

where  $M$  is defined as in Theorem 9,  $K_1 = 6 \cdot \frac{1 + \frac{1 + \kappa^2 d_{nz}^2 / \mu^2}{(1 - \kappa d_{nz} / \mu)^2}}{(1 - \kappa^2 \frac{d_z d_{nz}}{\mu \nu})^2}$ ,  $K_2 = 6 \left(1 - \kappa^2 \frac{d_z d_{nz}}{\mu \nu}\right)^{-2}$  in which we assume  $\kappa^2 \leq \frac{\mu \nu}{d_z d_{nz}}$  and  $\mu \geq \kappa d_{nz}$ .

In the orthogonal case, we can show the oracle inequalities [20] hold.

**Theorem 12** Suppose  $\Theta$  satisfies the sandwiching condition (4.16) and  $\mathbf{X}^T \mathbf{X} = \mathbf{I}$ . Then

$$E \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2^2 \leq (1 + \tau^2) \cdot \sum_1^n \min \left( \frac{2\varphi(\tau)}{\tau} \sigma^2 + \beta_i^2, \sigma^2 \right) \quad (4.25)$$

for any  $\tau > 1$ . Consequently, when  $\tau = \sqrt{2 \log n}$ ,

$$E \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2^2 \leq (2 \log n + 1) \left( \frac{\sigma^2}{\sqrt{\pi \log n}} + \sum \min(\beta_i^2, \sigma^2) \right) \quad (4.26)$$

for any  $n \geq 2$ .

This nonasymptotic result covers the soft-, the hard-, and the SCAD-thresholdings. It coincides with the classical soft-thresholding studies [20] and is sharper than [5, 61].

## 4.4 TISP Designs and Numerical Examples

### 4.4.1 An empirical study of TISPs

In this section, we demonstrate the empirical performance of TISPs by some simulation data. Although there are rich choices about  $\Theta$  in (4.9), we focus on three basic TISPs only in this subsection. In addition to the Soft-TISP, i.e., the lasso, we implemented Hard-TISP and SCAD-TISP, the thresholdings of which belong to the *hard-thresholding family*. The parameter  $a$  in SCAD-thresholding takes the default value, 3.7, based on a Bayesian argument [26]. As seen from the theoretical studies in Section 4.3, the last two should perform better than the lasso in variable selection. In generating the solution path for a grid of  $\lambda$ -values, we always set the initial point,  $\boldsymbol{\beta}^{(0)}$ , to be zero in Hard- or SCAD-TISP. (Note that a pathwise algorithm with warm start, which takes the previous estimate

associated with the old value of  $\lambda$  as the initial point of the procedure for the current value of  $\lambda$ , may be inappropriate for TISPs because it leads to bad solutions when nonconvex penalties are used.)

For comparison, the one-step LLA method, proposed by Zou and Li [63] for penalized likelihood models, is also included in our tests. They showed good asymptotics about one-step SCAD when  $n \rightarrow \infty$  and  $p$  is fixed, and demonstrated its performance in various numerical examples. The one-step LLA is actually a *weighted lasso* [45] with weights constructed from the OLS estimate using different penalty functions. According to our general result of weights in sparse regression (see Chapter 3 or [46]), it can achieve better sign consistency than the lasso as  $n$  grows to infinity. We are greatly interested in drawing a comparison between TISP and LLA since TISP also successfully solves the penalized regression problems.

We did experiments on two simulation datasets. Each dataset contains training data, validation data, and test data. We use  $\# = \text{“} \cdot / \cdot / \cdot \text{”}$  to denote the number of observations in the training data, validation data, and test data. Let  $\Sigma$  be the correlation matrix in generating  $\mathbf{X}$ , i.e., each row of  $\mathbf{X}$  is independently drawn from  $N(\mathbf{0}, \Sigma)$ . We use  $(\{a_1\}^{n_1}, \dots, \{a_k\}^{n_k})$  to denote the column vector made by  $n_1$   $a_1$ 's,  $\dots$ ,  $n_k$   $a_k$ 's consecutively in the following examples.

**Example 1.**  $\# = 20/100/200$ ,  $d = 8$ ,  $\beta = (\{3\}^1, \{1.5\}^1, \{0\}^2, \{2\}^1, \{0\}^3)$ ,  $\Sigma_{ij} = \rho^{|i-j|}$  with  $\rho = 0.5$ ,  $\sigma = 2, 3, 5, 8$ ; the corresponding signal-to-noise variance ratio ( $\beta^T \Sigma \beta / \sigma^2$ ) is 5.31, 2.36, 0.85, and 0.33, respectively.

**Example 2.**  $\# = 20/100/200$ ,  $d = 8$ ,  $\beta = (\{3\}^1, \{1.5\}^1, \{0\}^2, \{2\}^1, \{0\}^3)$ ,  $\Sigma_{ij} = \rho^{|i-j|}$  with  $\rho = 0.85$ ,  $\sigma = 2, 3, 5, 8$ ; the corresponding signal-to-noise variance ratio is 8.21, 3.65, 1.31, and 0.51, respectively.

Before an algorithm is applied, the columns of a regression matrix are all normalized to have a squared  $l_2$ -norm equal to the number of the observations; no centering is performed in these examples.

Each model is simulated 50 times, then, we measure the performance of each algorithm mainly by test error and sparsity error. The test error is characterized by the 40% trimmed-mean<sup>3</sup> of the scaled MSE (SMSE) on the test data, where SMSE is  $100 \cdot (\sum_{i=1}^N (\hat{y}_i - y_i)^2 / (N\sigma^2) - 1)$  defined for the test data. The sparsity error here is defined by the 40%

<sup>3</sup>Medians of errors are mostly used [50, 62] to measure the performance from multiple runs, but are not so stable for comparisons based on our experience. Discarding 20 highest and 20 lowest errors, we compute the average of the remaining 10.

trimmed-mean of the following 50 percentages:  $100 \cdot |\{i : \text{sgn}(\hat{\beta}_i) \neq \text{sgn}(\beta_i)\}|/d$ , which represents the number of inconsistent signs for each estimate compared to the true  $\beta$ . We also summarized the proper zero percentages,  $100\% \cdot |\{i : \beta_i = 0, \hat{\beta}_i = 0\}|/|\{i : \beta_i = 0\}|$ , and the proper nonzero percentages,  $100\% \cdot |\{i : \beta_i \neq 0, \hat{\beta}_i \neq 0\}|/|\{i : \beta_i \neq 0\}|$  in the table as follows.

First, although Zou and Li's one-step SCAD brings more sparsity than the lasso estimate (seen from the proper-sparsity and proper-nonsparsity), it is the worst in terms of test error. This is because the one-step SCAD is indeed a weighted lasso method and the OLS estimate used for weight construction may not be trustworthy, if, say, there is large noise, or high correlation between some variables. This phenomenon is serious in Example 2 where the OLS estimate can be unstable and misleading. Our Hard-TISP and SCAD-TISP clearly showed the remarkable *parsimoniousness* brought by nonconvex penalties. Instead of solving a  $l_1$ -constrained convex approximation as in the LLA method, our TISPs directly tackled the original (nonconvexly) penalized regressions and demonstrated better performance in both test-error and sparsity-error. (In fact, we *doubt* if the  $l_1$ -based one-step SCAD is truly able to solve the SCAD penalized regression, seen from the convex approximation in its derivation, and after comparing its estimate to the SCAD-TISP.) Hard-TISP and SCAD-TISP do not differ much here, which verifies the previous theoretical results regarding the hard-thresholding family in Section 4.3.

Hard-TISP and SCAD-TISP achieve smaller test error than the lasso which may introduce extra bias when the signal-to-noise ratio is medium or high. But when the noise level is very high, the lasso (Soft-TISP) yields a more accurate estimate than the two. This is in fact not so surprising. When the noise is relatively large compared to the signal, it is also necessary to shrink the nonzero coefficients even if the true ones are far from zero. In either hard- or SCAD-thresholding, there is basically no shrinkage offered for large nonzero coefficients, while the lasso does this by soft-thresholding (although the shrinkage amount is the same as the thresholding value). Fortunately, TISP still gives us good selection results and achieves parsimonious models. We can apply, for example, a second-time shrinkage to the coefficients of the selected variables. Of course, a better strategy is to take into account these two concerns – selection and shrinkage – *simultaneously* in building a model as probed in the next subsection.

		Lasso	One-step SCAD	Hard- TISP	SCAD- TISP	eNet	Hybrid- TISP
EX1, $\sigma = 2$	Test-err	<b>28.6</b>	<b>25.3</b>	<b>21.7</b>	<b>18.2</b>	<b>25.4</b>	<b>16.9</b>
	Spar-err	<b>31.8</b>	<b>12.5</b>	<b>0</b>	<b>12.5</b>	<b>31.0</b>	<b>0</b>
	<i>Prop-Z</i>	50.8%	91.2%	100%	89.5%	51.2%	100%
	<i>Prop-NZ</i>	100%	100%	100%	100%	100%	100%
EX1, $\sigma = 3$	Test-err	<b>27.8</b>	<b>27.3</b>	<b>25.9</b>	<b>25.8</b>	<b>23.4</b>	<b>18.4</b>
	Spar-err	<b>30.7</b>	<b>16.7</b>	<b>5.5</b>	<b>12.5</b>	<b>31.5</b>	<b>4.8</b>
	<i>Prop-Z</i>	50.8%	80.0%	93.2%	92.0%	47.3%	93.2%
	<i>Prop-NZ</i>	100%	87.2%	100%	100%	100.0%	100.0%
EX1, $\sigma = 5$	Test-err	<b>23.0</b>	<b>27.0</b>	<b>22.3</b>	<b>25.7</b>	<b>18.4</b>	<b>18.2</b>
	Spar-err	<b>32.0</b>	<b>25.0</b>	<b>12.5</b>	<b>25.0</b>	<b>31.5</b>	<b>17.4</b>
	<i>Prop-Z</i>	50.4%	80.0%	91.6%	80.0%	48.6%	91.9%
	<i>Prop-NZ</i>	86.2%	66.7%	85.1%	66.7%	100.0%	88.1%
EX1, $\sigma = 8$	Test-err	<b>15.4</b>	<b>20.4</b>	<b>20.3</b>	<b>17.1</b>	<b>14.1</b>	<b>11.7</b>
	Spar-err	<b>31.3</b>	<b>30.5</b>	<b>25.0</b>	<b>32.8</b>	<b>37.5</b>	<b>30.1</b>
	<i>Prop-Z</i>	72.3%	80.0%	94.9%	80.0%	70.4%	94.0%
	<i>Prop-NZ</i>	66.7%	33.3%	49.5%	33.3%	66.7%	66.7%
EX2, $\sigma = 2$	Test-err	<b>24.1</b>	<b>28.5</b>	<b>19.9</b>	<b>20.8</b>	<b>19.4</b>	<b>15.7</b>
	Spar-err	<b>31.0</b>	<b>32.0</b>	<b>12.5</b>	<b>12.5</b>	<b>30.8</b>	<b>12.5</b>
	<i>Prop-Z</i>	60.0%	74.5%	80.0%	90.5%	52.3%	80.0%
	<i>Prop-NZ</i>	100%	84.1%	100%	100%	100.0%	100.0%
EX2, $\sigma = 3$	Test-err	<b>19.9</b>	<b>29.1</b>	<b>19.8</b>	<b>20.2</b>	<b>14.3</b>	<b>14.0</b>
	Spar-err	<b>30.7</b>	<b>29.8</b>	<b>16.1</b>	<b>16.3</b>	<b>28.6</b>	<b>16.5</b>
	<i>Prop-Z</i>	60.0%	80.0%	91.1%	91.9%	55.8%	80.0%
	<i>Prop-NZ</i>	85.7%	66.7%	83.7%	66.7 %	100.0%	100.0 %
EX2, $\sigma = 5$	Test-err	<b>13.9</b>	<b>24.5</b>	<b>16.6</b>	<b>17.4</b>	<b>9.7</b>	<b>9.5</b>
	Spar-err	<b>31.0</b>	<b>31.3</b>	<b>25.0</b>	<b>25.0</b>	<b>30.0</b>	<b>25.0</b>
	<i>Prop-Z</i>	68.8%	80.0%	80.0%	80.0%	52.9%	73.3%
	<i>Prop-NZ</i>	66.7%	48.2%	66.7%	66.7%	100.0%	87.2%
EX2, $\sigma = 8$	Test-err	<b>10.4</b>	<b>18.2</b>	<b>13.8</b>	<b>15.6</b>	<b>7.2</b>	<b>6.8</b>
	Spar-err	<b>32.0</b>	<b>36.4</b>	<b>31.0</b>	<b>29.0</b>	<b>37.5</b>	<b>31.3</b>
	<i>Prop-Z</i>	71.0%	80.0%	92.1%	91.0%	49.3%	73.1%
	<i>Prop-NZ</i>	66.7%	33.3%	47.4%	33.3%	83.0%	66.7%

Table 4.1: Performance comparisons on the simulation data, in terms of test error, sparsity error, proper sparsity, and proper nonsparsity – all the numbers are 40% trimmed-mean of the 50 simulations. Six methods are listed here: lasso (Soft-TISP), one-step SCAD, Hard-TISP, SCAD-TISP, elastic net (eNet), and Hybrid-TISP.

#### 4.4.2 Hybrid-TISP for model selection and shrinkage

To deal with the low SNR problem, a promising approach is to modify the thresholding in Hard-TISP to include shrinkage for nonzero coefficients. Motivated by the thresholding function of ridge regression given by (4.10), we propose the following *hybrid*-thresholding:

$$\Theta(t; \lambda, \eta) = \begin{cases} 0, & \text{if } |t| < \lambda \\ \frac{t}{1+\eta}, & \text{if } |t| \geq \lambda \end{cases}. \quad (4.27)$$

The penalty constructed via the mechanism introduced in Subsection 4.2.1 is made up of two quadratic parts:

$$P(\theta; \lambda, \eta) = \begin{cases} -\frac{1}{2}\theta^2 + \lambda|\theta|, & \text{if } |\theta| < \frac{\lambda}{1+\eta} \\ \frac{1}{2}\eta\theta^2 + \frac{1}{2}\frac{\lambda^2}{1+\eta}, & \text{if } |\theta| \geq \frac{\lambda}{1+\eta} \end{cases}. \quad (4.28)$$

We have seen the first quadratic part in the smooth hard-penalty (which leads to the same solution as the discrete  $l_0$ -penalty); the second part resembles a ridge penalty. See Figure 4.1 below. Simple calculations show that this  $P$  satisfies the BCC (cf. (4.11)) with  $\mathbf{H} = \mathbf{I}$ , and thus Theorem 8 holds. We can apply (4.14) given an arbitrary design matrix. The corresponding TISP (referred to as **Hybrid-TISP**) converges.

Moreover, we have the following nonasymptotic result in parallel to Theorem 10. Recall that  $k_0 = \|\mathbf{X}\|_2$ ,  $\Sigma^{(s)} = \Sigma/n = \mathbf{X}^T \mathbf{X}/n$ ,  $\mu = \mu_{\min}(\Sigma_{nz,nz}^{(s)})$  and  $\kappa \triangleq \max_{i \in z} \|\Sigma_{i,nz}^{(s)}\|_2 / \sqrt{d_{nz}}$ . Define  $\iota \triangleq \min |(\Sigma + \eta \mathbf{I})^{-1} \Sigma \beta|$ , the minimum absolute value in the partial ridge estimate without noise. Let  $p_e$  be the probability of Hybrid-TISP estimates having inconsistent zeros, that is, for any  $\hat{\beta}$ , there exists some  $i$  or  $j$  such that  $\hat{\beta}_{z,i} \neq 0$  or  $\hat{\beta}_{nz,j} = 0$ .

**Theorem 13** *Assume  $\mu > 0$ , and  $\lambda, \eta$  are chosen such that  $\kappa \leq \frac{\lambda}{\|\beta_{nz}\|_2 \sqrt{d_{nz}}} \frac{n\mu + \eta}{n\eta}$  and  $\iota \geq \frac{\lambda}{k_0^2 + \eta}$ . Then*

$$p_e \leq 2d_z \varphi(M'')/M'' + 2d_{nz} \varphi(L'')/L'', \quad (4.29)$$

where  $M'' = \frac{1}{\sqrt{n}\sigma} \left( \lambda - \frac{n\eta}{n\mu + \eta} \kappa \|\beta_{nz}\|_2 \sqrt{d_{nz}} \right)$ ,  $L'' = \frac{n\mu + \eta}{\sqrt{n\mu}\sigma} \left( \iota - \frac{\lambda}{k_0^2 + \eta} \right)$ .

Hybrid-TISP successfully offers both selection and shrinkage in estimating  $\beta$ . Before going into the numerical results, we summarize the traits of the design of Hybrid-TISP as follows. (a) Its penalty provides us a trade-off between the  $l_0$ -penalty and the  $l_2$ -penalty

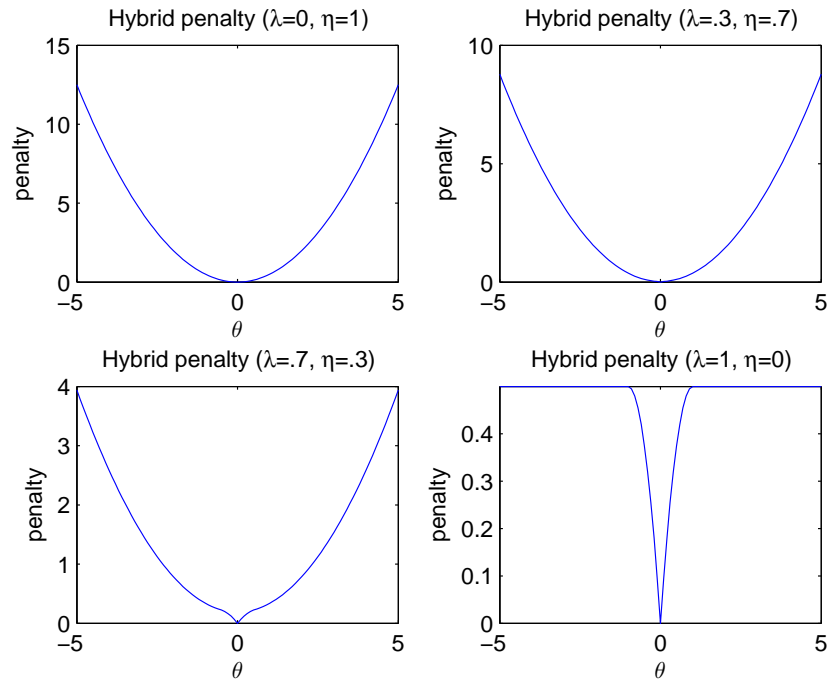


Figure 4.1: The penalty defined by hybrid-thresholding. As  $\lambda$  and  $\eta$  vary, it takes the smooth hard-penalty and the ridge penalty as extremes.

(ridge-penalty), and takes the two as extremes, from which we secure selection and shrinkage simultaneously. In particular, the selection is achieved by a penalty more like  $l_0$  than  $l_1$ , seen from the penalty function, or the iterative thresholding. (b) Hybrid-TISP avoids double shrinkage. Double shrinkage is a serious problem in the design of naive elastic net [62] which simply adopts a linear combination of the  $l_1$ -penalty and the  $l_2$ -penalty. However, the  $l_1$ -penalty also plays a role in shrinking the nonzero coefficients in addition to the  $l_2$ -penalty. By contrast, Hybrid-TISP deals with the zeros and the nonzeros separately, by hard-thresholding and ridge-thresholding, respectively; there is no overlapping between them. (c) We have two parameters,  $\lambda$  and  $\eta$ , responsible for selection and shrinkage respectively. One drawback of the lasso is that it uses the same parameter to control both selection and shrinkage [36]. Therefore, it may result in insufficient zeros even if the SNR is pretty high, as shown clearly in Table 4.1. Hybrid-TISP has  $\lambda, \eta$  designed for the two different purposes. (d) The TISP selecting and shrinking *interplay* with each other during the iteration till in the end we successfully achieve selection/shrinkage balance in the final

estimate. This is in contrast to the relaxed lasso [36] which treats selection and shrinkage as separate steps in building a model. (e) Finally, Hybrid-TISP is a very simple procedure to implement.

In the implementation of Hybrid-TISP, an empirical parameter search is usually needed to determine the values for  $(\lambda, \eta)$ . We adopted the *alternative* search strategy [45] which has been shown to be fast and efficacious: First, letting  $\eta = \lambda/4$ , generate the solution path for different values of  $\lambda$  and search over the path to get a solution with the smallest validation error at, say,  $(\lambda^{(o)}, \eta^{(o)})$ . Next, fixing  $\eta$  at  $\eta^{(o)}$ , search along the  $\lambda$ -path to get an optimal value for  $\lambda$ , denoted by  $\lambda^{(oo)}$ , the corresponding solution having the smallest validation error. Then, with  $\lambda$  fixed at  $\lambda^{(oo)}$ , we do the last search over the  $\eta$ -path. Finally, compare the results from the 3 searches, and take  $(\lambda, \eta)$  to be the one minimizing the validation error. The results are reported in Table 4.1. We also included the elastic net (eNet) in the experiments, which has two regularization parameters as well. Note that we have to generate and search along 6 solution paths in the eNet to tune the parameters [62].

Seen from the table, Hybrid-TISP has amazing performance in both accuracy and sparsity. In fact, it beats all the other methods in all situations, a phenomenon rarely seen in empirical studies. We briefly summarize the story as follows. When the noise level is low or medium, the value of  $\lambda$  in the lasso is limited by the amount of shrinkage and thus gives insufficient sparsity. Large noise alleviates the problem but there is still much room for the improvement of test-error and sparsity-error because the amount of shrinkage may not equal to the thresholding value in the selection. The weighted lasso like the one-step SCAD has very limited power because the OLS estimate may be inaccurate and misleading for weight construction. Benefiting from the  $l_2$ -penalty, the eNet shows much better accuracy in the case of large noise and/or high correlation between the variables; nevertheless, the sparsity of the estimate can be *seriously* hurt when the ridge penalty must take control. And it seems possible to improve its test-error further by incorporating this sparsity in estimation. All of these problems can be resolved by Hybrid-TISP, which achieves the right balance between shrinkage and selection. Its test error is consistently lower than the eNet, and more importantly, Hybrid-TISP provides a parsimonious model as Hard-TISP.

## 4.5 Discussion

We have proposed the thresholding-based iterative selection procedures for solving nonconvexly penalized regressions. In fact, people have long before noticed the weakness of the convex  $l_1$ -constraint (or the soft-thresholding) in wavelets and have designed many different forms of nonconvex penalties to increase model sparsity and accuracy. But for a nonorthogonal regression matrix, there is great difficulty in both investigating the performance in theory and solving the problem in computation. TISP provides a simple and efficient way to tackle this.

Somewhat different than other studies, we started from thresholding rules rather than penalty functions. Indeed, there is a universal connection between them. But a drawback of the latter is its non-unique form: different penalties may result in the same estimate and the same thresholding. Moreover, starting from  $\Theta$  greatly facilitates the computation and the analysis. In fact, some penalty designs may even have a better explanation from  $\Theta$ , or equivalently, Huber's  $\psi$ -function (4.7) — for example, the SCAD-penalty (recall that it is defined by its derivative) seems to originate from Hampel's three-part redescending  $\psi$  [29, 4].

Using a thresholding rule in the hard-thresholding family, TISP gives good selection results. Our novel Hybrid-TISP, accomplishing a fusion between  $l_0$ -penalty and  $l_2$ -penalty based on the hard-thresholding and the ridge thresholding, shows superior performance and beats the commonly used methods in both test-error and sparsity. It is worth mentioning that in contrast to [30, 5, 4], where more than one tuning parameter is considered a drawback and unnecessary, we believe a good procedure should have two explicit regularization parameters to control and balance selection and shrinkage.

We assume the penalty function  $P$  is dependent on  $\beta$  and  $\lambda$  only. Therefore the iterative weighting, substituting the nonnegative garrote [30] for  $\Theta$  in TISP, is not covered by the studies in this chapter. In fact, with  $\beta$  involved in  $P$ , it might be difficult to optimize in the second step of the mechanism introduced in Section 4.1.

The solution path associated with a nonconvex penalty is generally not continuous in  $\lambda$ . For example, even for the transformed  $L_1$  penalty [31] which is differentiable to any order on  $(0, +\infty)$ , the solution path still has no  $\lambda$ -continuity practically. Hence a pathwise algorithm is not appropriate here. Empirically, using a zero estimate as the start in nonconvex TISPs works pretty well.

The generalization of TISP to GLM seems straightforward; we will investigate this topic in the next paper. Other future studies include developing some acceleration techniques for TISP and deriving some risk oracles in theory. Finally, TISP fits perfectly into the Accelerated Annealing (see Chapter 3 or [46]) and it is very promising to use this technique to solve the generic sparse regression, such as the supervised clustering problem [46].

## 4.6 Proofs of Theorem 8, Proposition 8, and Proposition 9

Let's consider the orthogonal case first. Define  $Q(\boldsymbol{\gamma}) = \|\boldsymbol{\gamma} - \boldsymbol{\alpha}\|_2^2/2 + P(\boldsymbol{\gamma}; \lambda)$ , where  $\boldsymbol{\alpha}$  is a known vector. Let  $\boldsymbol{\gamma}_o = \arg \min Q(\boldsymbol{\gamma})$ . By the construction of  $P$  and Proposition 7,  $\boldsymbol{\gamma}_o$  satisfies  $\boldsymbol{\gamma}_o - \boldsymbol{\alpha} + s(\boldsymbol{\gamma}_o; \lambda) = \mathbf{0}$ .

$$\begin{aligned} Q(\boldsymbol{\gamma}_o + \mathbf{h}) - Q(\boldsymbol{\gamma}_o) &= \frac{1}{2}\|\boldsymbol{\gamma}_o + \mathbf{h} - \boldsymbol{\alpha}\|_2^2 - \frac{1}{2}\|\boldsymbol{\gamma}_o - \boldsymbol{\alpha}\|_2^2 + P(\boldsymbol{\gamma}_o + \mathbf{h}; \lambda) - P(\boldsymbol{\gamma}_o; \lambda) \\ &= \frac{1}{2}\|\mathbf{h}\|_2^2 + \langle \mathbf{h}, \boldsymbol{\gamma}_o - \boldsymbol{\alpha} \rangle + P(\boldsymbol{\gamma}_o + \mathbf{h}; \lambda) - P(\boldsymbol{\gamma}_o; \lambda) \\ &= \frac{1}{2}\|\mathbf{h}\|_2^2 + (P(\boldsymbol{\gamma}_o + \mathbf{h}; \lambda) - P(\boldsymbol{\gamma}_o; \lambda) - \langle \mathbf{h}, \mathbf{s} \rangle) \\ &\geq \frac{1}{2}\|\mathbf{h}\|_2^2 - \frac{1}{2}\mathbf{h}^T \mathbf{H} \mathbf{h} = \frac{1}{2}\mathbf{h}^T (\mathbf{I} - \mathbf{H}) \mathbf{h}. \end{aligned}$$

This inequality is due to the BCC (4.11). On the other hand, we know

$$Q(\boldsymbol{\gamma}_o + \mathbf{h}) - Q(\boldsymbol{\gamma}_o) \geq 0.$$

In summary, we get

$$Q(\boldsymbol{\gamma}_o + \mathbf{h}) - Q(\boldsymbol{\gamma}_o) \geq \frac{1}{2}\mathbf{h}^T \mathbf{A} \mathbf{h}, \quad (4.30)$$

for both  $\mathbf{A} = \mathbf{I} - \mathbf{H}$  and  $\mathbf{A} = \mathbf{0}$ ; formally, we write  $\mathbf{A} = (\mathbf{I} - \mathbf{H}) \vee \mathbf{0}$ . Note that (4.30) is a global result for *any*  $\mathbf{h}$ .

Now look at the TISP. For convenience, redefine the  $g$  in (4.2) as

$$g(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \frac{1}{2}\|\mathbf{X}\boldsymbol{\gamma} - \mathbf{y}\|_2^2 + P(\boldsymbol{\gamma}; \lambda) + \frac{1}{2}(\boldsymbol{\gamma} - \boldsymbol{\beta})^T (\mathbf{I} - \boldsymbol{\Sigma})(\boldsymbol{\gamma} - \boldsymbol{\beta}).$$

Then given  $\boldsymbol{\beta}$ , we can write  $g$  as

$$g(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \frac{1}{2} \|\boldsymbol{\gamma} - ((\mathbf{I} - \boldsymbol{\Sigma})\boldsymbol{\beta} + \mathbf{X}^T \mathbf{y})\|_2^2 + C(\mathbf{X}, \mathbf{y}, \boldsymbol{\beta}),$$

and apply (4.30) with  $\boldsymbol{\alpha} = (\mathbf{I} - \boldsymbol{\Sigma})\boldsymbol{\beta} + \mathbf{X}^T \mathbf{y}$ ,

$$g(\boldsymbol{\beta}, \boldsymbol{\gamma}_o(\boldsymbol{\beta}) + \mathbf{h}) - g(\boldsymbol{\beta}, \boldsymbol{\gamma}_o(\boldsymbol{\beta})) \geq \frac{1}{2} \mathbf{h}^T ((\mathbf{I} - \mathbf{H}) \vee \mathbf{0}) \mathbf{h}, \quad \forall \mathbf{h} \quad (4.31)$$

Correspondingly, for the TISP iterates  $\boldsymbol{\beta}^{(j)}$ , we have

$$\begin{aligned} & f(\boldsymbol{\beta}^{(j+1)}) + \frac{1}{2} (\boldsymbol{\beta}^{(j+1)} - \boldsymbol{\beta}^{(j)})^T (\mathbf{I} - \boldsymbol{\Sigma}) (\boldsymbol{\beta}^{(j+1)} - \boldsymbol{\beta}^{(j)}) = g(\boldsymbol{\beta}^{(j)}, \boldsymbol{\beta}^{(j+1)}) \\ & \leq g(\boldsymbol{\beta}^{(j)}, \boldsymbol{\beta}^{(j)}) - \frac{1}{2} (\boldsymbol{\beta}^{(j+1)} - \boldsymbol{\beta}^{(j)})^T ((\mathbf{I} - \mathbf{H}) \vee \mathbf{0}) (\boldsymbol{\beta}^{(j+1)} - \boldsymbol{\beta}^{(j)}) \\ & = f(\boldsymbol{\beta}^{(j)}) - \frac{1}{2} (\boldsymbol{\beta}^{(j+1)} - \boldsymbol{\beta}^{(j)})^T ((\mathbf{I} - \mathbf{H}) \vee \mathbf{0}) (\boldsymbol{\beta}^{(j+1)} - \boldsymbol{\beta}^{(j)}). \end{aligned}$$

That is,

$$f(\boldsymbol{\beta}^{(j)}) - f(\boldsymbol{\beta}^{(j+1)}) \geq \frac{1}{2} (\boldsymbol{\beta}^{(j+1)} - \boldsymbol{\beta}^{(j)})^T ((\mathbf{I} - \mathbf{H}) \vee \mathbf{0} + \mathbf{I} - \boldsymbol{\Sigma}) (\boldsymbol{\beta}^{(j+1)} - \boldsymbol{\beta}^{(j)}). \quad (4.32)$$

Now (4.12) and (4.13) can be obtained after simple calculations.

As for Proposition 8, let  $\boldsymbol{\beta}^{(j_k)} \rightarrow \boldsymbol{\beta}^*$  as  $k \rightarrow \infty$ . Under the condition  $\mu_{\max}(\boldsymbol{\Sigma}) < 1 \vee (2 - \mu_{\max}(\mathbf{H}))$ , Theorem 8 states that

$$\|\boldsymbol{\beta}^{(j_{k+1})} - \boldsymbol{\beta}^{(j_k)}\|_2^2 \leq (f(\boldsymbol{\beta}^{(j_k)}) - f(\boldsymbol{\beta}^{(j_{k+1})})) / C \leq (f(\boldsymbol{\beta}^{(j_k)}) - f(\boldsymbol{\beta}^{(j_{k+1})})) / C \rightarrow 0.$$

That is,  $\Theta((\mathbf{I} - \boldsymbol{\Sigma})\boldsymbol{\beta}^{(j_k)} + \mathbf{X}^T \mathbf{y}; \lambda) - \boldsymbol{\beta}^{(j_k)} \rightarrow 0$ . Therefore,  $\boldsymbol{\beta}^*$  is a fixed point of TISP.

Finally, we prove Proposition 9. Noticing that  $\boldsymbol{\gamma}_o(\boldsymbol{\beta}^*) = \boldsymbol{\beta}^*$ , we get the following inequality from (4.31)

$$g(\boldsymbol{\beta}^*, \boldsymbol{\beta}^* + \mathbf{h}) - g(\boldsymbol{\beta}^*, \boldsymbol{\beta}^*) \geq \frac{1}{2} \mathbf{h}^T ((\mathbf{I} - \mathbf{H}) \vee \mathbf{0}) \mathbf{h}, \quad \forall \mathbf{h}.$$

Since  $g(\boldsymbol{\beta}^*, \boldsymbol{\beta}^*) = f(\boldsymbol{\beta}^*)$ ,

$$\begin{aligned} f(\boldsymbol{\beta}^* + \mathbf{h}) + \frac{1}{2}\mathbf{h}^T(\mathbf{I} - \boldsymbol{\Sigma})\mathbf{h} &\geq f(\boldsymbol{\beta}^*) + \frac{1}{2}\mathbf{h}^T((\mathbf{I} - \mathbf{H}) \vee \mathbf{0})\mathbf{h}, \quad \forall \mathbf{h} \\ \Rightarrow f(\boldsymbol{\beta}^* + \mathbf{h}) - f(\boldsymbol{\beta}^*) &\geq \frac{1}{2}\mathbf{h}^T((\mathbf{I} - \mathbf{H}) \vee \mathbf{0} + \boldsymbol{\Sigma} - \mathbf{I})\mathbf{h}, \quad \forall \mathbf{h}. \end{aligned}$$

Therefore, if  $\mu(\boldsymbol{\Sigma}) \geq \mu_{\max}(\mathbf{H})$ ,  $\boldsymbol{\beta}^*$  is a global minimizer of  $f$ .  $\blacksquare$

## 4.7 Proof Outlines of Theorem 9, Theorem 10, Theorem 11, and Theorem 12

These theorems have all been essentially proved in Chapter 2 (or see a previous report [45]), using the generalized sign form of the KKT equation (4.17). First, the proof of Theorem 2 applies to a general TISP due to Proposition 10 and so Theorem 9 and Theorem 11 are true. For Theorem 10, noticing that (a)  $\widetilde{\text{sgn}}(u) = 0, \forall |u| > c\tau$  by definition and (b)  $p_s \geq P\left(\hat{\boldsymbol{\beta}}_z^{(s)} = 0, \text{ and } |\hat{\boldsymbol{\beta}}_{nz}^{(s)}| > c\tau^{(s)}\right)$  with  $\boldsymbol{\beta}^{(s)} = \boldsymbol{\beta}\sqrt{n}$ ,  $\tau^{(s)} = \tau/\sqrt{n}$ , we can prove it following the same lines as the proof of the sign consistency part in Theorem 2.

Finally, we outline the proof of Theorem 12. Let  $\hat{\boldsymbol{\beta}}^H, \hat{\boldsymbol{\beta}}^S$  denote the hard- and soft-thresholding estimates with threshold value  $\tau$ . It is easy to see  $\hat{\boldsymbol{\beta}}^H, \hat{\boldsymbol{\beta}}^S$ , and  $\hat{\boldsymbol{\beta}}$  all have the same sign and  $\hat{\boldsymbol{\beta}}$  is sandwiched by the other two. Therefore,  $E\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2^2 \leq \sum E(\max((\hat{\beta}_i^S - \beta_i)^2, (\hat{\beta}_i^H - \beta_i)^2))$ . In the proof of Theorem 3, we showed for  $y = \mu + \epsilon$  (all are scalars) with  $\epsilon \sim N(0, 1)$ , for both soft-thresholding and hard-thresholding, the risk function  $\rho(\tau, \mu)$  is bounded:  $\rho(\tau, \mu) \leq 1 + \tau^2$  for  $\tau > 1$ , and  $\rho(\tau, \mu) \leq \rho(\tau, 0) + 1.2\mu^2$ . These bounds directly lead to the oracle inequalities given by Theorem 12.

## 4.8 Proof of Theorem 13

In this section, all inequalities and the absolute value ‘ $|\cdot|$ ’ are understood in the component-wise sense.

First we calculate the generalized sign for the hybrid-thresholding (4.27)

$$\widetilde{\text{sgn}}(u; \lambda, \eta) = \begin{cases} \in [-1, 1], & \text{if } u = 0 \\ 0, & \text{if } |u| \in (0, \frac{\lambda}{1+\eta}) \\ \frac{\eta}{\lambda} \cdot u, & \text{if } |u| \geq \frac{\lambda}{1+\eta} \end{cases}. \quad (4.33)$$

And note that  $\tau(\lambda) = \lambda$ . The KKT equation for the Hybrid-TISP estimate  $\hat{\boldsymbol{\beta}}$  from (4.14) is

$$\boldsymbol{\Sigma}\hat{\boldsymbol{\beta}} = \mathbf{X}^T \mathbf{y} - \lambda \widetilde{\text{sgn}} \left( \hat{\boldsymbol{\beta}}; \frac{\lambda}{k_0^2}, \frac{\eta}{k_0^2} \right), \quad (4.34)$$

where  $k_0 = \|\mathbf{X}\|_2$ .

The proof still follows the lines of the proof for Theorem 2 (or see [45]). Assume, for the moment,  $\mathbf{X}$  has been column-normalized such that the diagonal entries of  $\boldsymbol{\Sigma} = \mathbf{X}^T \mathbf{X}$  are all 1. Clearly,  $\hat{\boldsymbol{\beta}}_z = \mathbf{0}$ ,  $|\hat{\boldsymbol{\beta}}_{nz}| \geq \frac{\lambda}{k_0^2 + \eta}$  is a sufficient condition for the zero consistency of  $\hat{\boldsymbol{\beta}}$ . From Lemma 1, the KKT equation is equivalent to

$$\begin{cases} \mathbf{S}_z \hat{\boldsymbol{\beta}}_z = (\mathbf{X}_z^T - \boldsymbol{\Sigma}_{z,nz} \boldsymbol{\Sigma}_{nz}^{-1} \mathbf{X}_{nz}^T) \boldsymbol{\epsilon} + \lambda \boldsymbol{\Sigma}_{z,nz} \boldsymbol{\Sigma}_{nz}^{-1} \widetilde{\text{sgn}}(\hat{\boldsymbol{\beta}}_{nz}) - \lambda \widetilde{\text{sgn}}(\hat{\boldsymbol{\beta}}_z) \\ \hat{\boldsymbol{\beta}}_{nz} = \boldsymbol{\beta}_{nz} + \boldsymbol{\Sigma}_{nz}^{-1} (\mathbf{X}_{nz}^T \boldsymbol{\epsilon} - \lambda \widetilde{\text{sgn}}(\hat{\boldsymbol{\beta}}_{nz})) - \boldsymbol{\Sigma}_{nz}^{-1} \boldsymbol{\Sigma}_{z,nz} \hat{\boldsymbol{\beta}}_z \end{cases}$$

Our calculations based on the definition of  $\widetilde{\text{sgn}}$  show that

$$\begin{cases} \lambda \widetilde{\text{sgn}}(\mathbf{0}) = \{ \mathbf{X}_z^T - \boldsymbol{\Sigma}_{z,nz} \boldsymbol{\Sigma}_{nz}^{-1} [\mathbf{I} - \eta(\boldsymbol{\Sigma}_{nz} + \eta \mathbf{I})^{-1}] \mathbf{X}_{nz}^T \} \boldsymbol{\epsilon} + \eta \boldsymbol{\Sigma}_{z,nz} (\boldsymbol{\Sigma}_{nz} + \eta \mathbf{I})^{-1} \boldsymbol{\beta}_{nz} \\ \hat{\boldsymbol{\beta}}_{nz} = (\boldsymbol{\Sigma}_{nz} + \eta \mathbf{I})^{-1} \boldsymbol{\Sigma}_{nz} \boldsymbol{\beta}_{nz} + (\boldsymbol{\Sigma}_{nz} + \eta \mathbf{I})^{-1} \mathbf{X}_{nz}^T \boldsymbol{\epsilon} \end{cases}$$

Define

$$\begin{aligned} A &\triangleq \{ |\{ \mathbf{X}_z^T - \boldsymbol{\Sigma}_{z,nz} \boldsymbol{\Sigma}_{nz}^{-1} [\mathbf{I} - \eta(\boldsymbol{\Sigma}_{nz} + \eta \mathbf{I})^{-1}] \mathbf{X}_{nz}^T \} \boldsymbol{\epsilon} + \eta \boldsymbol{\Sigma}_{z,nz} (\boldsymbol{\Sigma}_{nz} + \eta \mathbf{I})^{-1} \boldsymbol{\beta}_{nz}| \leq \lambda \} \\ V &\triangleq \left\{ |(\boldsymbol{\Sigma}_{nz} + \eta \mathbf{I})^{-1} \boldsymbol{\Sigma}_{nz} \boldsymbol{\beta}_{nz} + (\boldsymbol{\Sigma}_{nz} + \eta \mathbf{I})^{-1} \mathbf{X}_{nz}^T \boldsymbol{\epsilon}| \geq \frac{\lambda}{k_0^2 + \eta} \right\}. \end{aligned}$$

Then  $p_e \leq P(A^c \cup V^c) \leq P(A^c) + P(V^c)$ .

To bound the first probability, noticing that

$$|\eta \boldsymbol{\Sigma}_{z,nz} (\boldsymbol{\Sigma}_{nz} + \eta \mathbf{I})^{-1} \boldsymbol{\beta}_{nz}| \leq \kappa \sqrt{d_{nz}} \frac{\eta}{\mu + \eta} \|\boldsymbol{\beta}_{nz}\|_2$$

we have

$$P(A^c) \leq P \left( \max |\boldsymbol{\epsilon}'_1| \geq \lambda - \kappa \sqrt{d_{nz}} \frac{\eta}{\mu + \eta} \|\boldsymbol{\beta}_{nz}\|_2 \right),$$

where  $\boldsymbol{\epsilon}'_1 = \{\mathbf{X}_z^T - \boldsymbol{\Sigma}_{z,nz}\boldsymbol{\Sigma}_{nz}^{-1}[\mathbf{I} - \eta(\boldsymbol{\Sigma}_{nz} + \eta\mathbf{I})^{-1}]\mathbf{X}_{nz}^T\}\boldsymbol{\epsilon}$ . Since

$$\text{var}(\boldsymbol{\epsilon}'_1) = \sigma^2 \{\boldsymbol{\Sigma}_z - \boldsymbol{\Sigma}_{z,nz}[\mathbf{I} - \eta^2(\boldsymbol{\Sigma}_{nz} + \eta\mathbf{I})^{-2}]\boldsymbol{\Sigma}_{nz}^{-1}\boldsymbol{\Sigma}_{z,nz}^T\},$$

$\text{diag}(\text{var}(\boldsymbol{\epsilon}'_1)) \leq \sigma^2 \text{diag}(\boldsymbol{\Sigma}_z) \leq \sigma^2 \mathbf{1}$ . It follows from Lemma 2 that

$$P(A^c) \leq P\left(\max |\boldsymbol{\epsilon}'_1| \sigma \geq \lambda - \kappa \sqrt{d_{nz}} \frac{\eta}{\mu + \eta} \|\boldsymbol{\beta}_{nz}\|_2\right),$$

where  $\boldsymbol{\epsilon}'_1 \sim N(\mathbf{0}, \mathbf{I}_{d_{nz} \times d_{nz}})$ . Define  $M'' = \frac{1}{\sigma} \left( \lambda - \kappa \frac{\eta}{\mu + \eta} \sqrt{d_{nz}} \|\boldsymbol{\beta}_{nz}\|_2 \right)$ . We obtain

$$P(A^c) \leq 2d_z \Phi([M'', +\infty)) \leq 2d_z \varphi(M'')/M''.$$

Next let's consider  $P(V^c)$ . Let  $\boldsymbol{\epsilon}'_2 = (\boldsymbol{\Sigma}_{nz} + \eta\mathbf{I})^{-1} \mathbf{X}_{nz}^T \boldsymbol{\epsilon}$ . Then

$$P(V^c) \leq P\left(\max |\boldsymbol{\epsilon}'_2| \geq \iota - \frac{\lambda}{k_0^2 + \eta}\right).$$

Since  $\text{var}(\boldsymbol{\epsilon}'_2) = (\boldsymbol{\Sigma}_{nz} + \eta\mathbf{I})^{-1} \boldsymbol{\Sigma}_{nz} (\boldsymbol{\Sigma}_{nz} + \eta\mathbf{I})^{-1} \sigma^2$ ,  $\text{diag}(\text{var}(\boldsymbol{\epsilon}'_2)) \leq \frac{\mu\sigma^2}{(\mu+\eta)^2}$ . By Lemma 2 again, we know

$$P(V^c) \leq P\left(\max |\boldsymbol{\epsilon}'_2| \geq \frac{\mu + \eta}{\sqrt{\mu}\sigma} \left( \iota - \frac{\lambda}{k_0^2 + \eta} \right)\right),$$

where  $\boldsymbol{\epsilon}'_2 \sim N(\mathbf{0}, \mathbf{I}_{d_{nz} \times d_{nz}})$ . Define  $L'' = \frac{\mu + \eta}{\sqrt{\mu}\sigma} \left( \iota - \frac{\lambda}{k_0^2 + \eta} \right)$ . It follows that

$$P(V^c) \leq 2d_{nz} \Phi([L'', +\infty)) \leq 2d_{nz} \varphi(L'')/L''.$$

We assumed  $\mathbf{x}_i^T \mathbf{x}_i = 1$   $i = 1, \dots, d$  in the above derivation. If the  $l_2$ -norm of each column of  $\mathbf{X}$  is no greater than  $\sigma_{\max}$ , it is not difficult to know that we only need to replace the  $\boldsymbol{\beta}$ ,  $\hat{\boldsymbol{\beta}}$ ,  $\lambda$ ,  $\eta$ , by  $\boldsymbol{\beta} \cdot \sigma_{\max}$ ,  $\hat{\boldsymbol{\beta}} \cdot \sigma_{\max}$ ,  $\lambda/\sigma_{\max}$ ,  $\eta/\sigma_{\max}^2$ , respectively. The proof of Theorem 13 is now complete if  $\sigma_{\max} = \sqrt{n}$ .  $\blacksquare$

# Bibliography

- [1] E. Amaldi and V. Kann. On the approximability of minimizing non zero variables or unsatisfied relations in linear systems. *Theoretical Computer Science*, 209:237–260, 1998.
- [2] T.W. Anderson. *The Statistical Analysis of Time Series*. Wiley, New York, 1971.
- [3] A. Antoniadis. Wavelets in statistics: a review (with discussion). *Italian Journal of Statistics*, 6:97–144, 1997.
- [4] A. Antoniadis. Wavelet methods in statistics: Some recent developments and their applications. *Statistics Surveys*, 1:16–55, 2007.
- [5] A. Antoniadis and J. Fan. Regularization of wavelets approximations. *JASA*, 96:939–967, 2001.
- [6] M. S. Bazaraa and C. M. Shetty. *Nonlinear Programming: Theory and Algorithms*. John Wiley & Sons, 1979.
- [7] D. Bertsekas. *Nonlinear Programming*. Athena Scientific, 1999.
- [8] P. J. Bickel, Y. Ritov, and A. B. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics*, 2008. To appear.
- [9] J. C. G. Boot. On sensitivity analysis in convex quadratic programming problems. *Operations Research*, 11(5):771–786, 1963.
- [10] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, Cambridge, MA, 2004.

- [11] L. Breiman. Better subset regression using the nonnegative garotte. *Technometrics*, 37:373–384, 1995.
- [12] F. E. Browder and W. V. Petryshyn. Construction of fixed points of nonlinear mappings in Hilbert space. *Journal of Mathematical Analysis and Applications*, 20(2):197–228, 1967.
- [13] F. Bunea, A. B. Tsybakov, and M.H. Wegkamp. Sparsity oracle inequalities for the lasso. *Electronic Journal of Statistics*, 1:169–194, 2007.
- [14] E. Candès. Modern statistical estimation via oracle inequalities. *Acta Numerica*, 15:257–325, 2006.
- [15] E. Candès, J. Romberg, and T. Tao. Stable signal recovery from incomplete and inaccurate measurements. *Comm. Pure Appl. Math.*, 59:1207–1223, 2006.
- [16] E. Candès and T. Tao. The Dantzig selector: statistical estimation when  $p$  is much smaller than  $n$ . *Annals of Statistics*, 35:2392–2404, 2005.
- [17] S. Chen, D. Donoho, and M. Saunders. Atomic decomposition by basis pursuit. *SIAM Journal of Scientific Computing*, 20(1):33–61, 1998.
- [18] I. Daubechies, M. Defrise, and C. De Mol. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Communications on Pure and Applied Mathematics*, 57:1413–1457, 2004.
- [19] D. Donoho, M. Elad, and V. Temlyakov. Stable recovery of sparse overcomplete representations in the presence of noise. *IEEE Transactions on Information Theory*, 52:6–18, 2006.
- [20] D. Donoho and I. Johnstone. Ideal spatial adaptation via wavelet shrinkages. *Biometrika*, 81:425–455, 1994.
- [21] W.G. Dotson Jr. On the Mann iterative process. *Trans. Amer. Math. Soc.*, 149:65–73, 1970.
- [22] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *Annals of Statistics*, 32:407–499, 2004.

- [23] B. Efron, T. Hastie, and R. Tibshirani. Discussion of ‘the Dantzig selector’ by e. candes and t. tao. *Annals of Statistics*, 35, 2007. 2358–2364.
- [24] H. W. Engl, M. Hanke, and A. Neubauer. *Regularization of Inverse Problems*. Kluwer, Dordrecht, 1996.
- [25] J. Fan. Comment on ‘Wavelets in Statistics: A Review’ by A. Antoniadis. *Italian Journal of Statistics*, 6:97–144, 1997.
- [26] J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.*, 96:1348–1360, 2001.
- [27] J. Friedman, T. Hastie, H. Hoffing, and Robert Tibshirani. Pathwise coordinate optimization. *Annals of Applied Statistics*, 1:302, 2007.
- [28] W. Fu. Penalized regressions: the bridge vs the lasso. *JCGS*, 7(3):397–416, 1998.
- [29] I. Gannaz. Robust estimation and wavelet thresholding in partial linear models. Technical report, University Joseph Fourier, Grenoble, France, 2006.
- [30] H.-Y. Gao. Wavelet shrinkage denoising using the non-negative garrote. *J. Comput. Graph. Statist.*, 7:469–488, 1998.
- [31] D. Geman and G. Reynolds. Constrained restoration and the recovery of discontinuities. *IEEE PAMI*, 14(3):367–383, 1992.
- [32] C. J. Geyer. On the asymptotics of convex stochastic optimization. 1996.
- [33] D. R. Hunter and K. Lange. Rejoinder to discussion of ‘Optimization transfer using surrogate objective functions’. *J. Comput. Graphical Stat*, 9:52–59, 2000.
- [34] K. Knight and W. Fu. Asymptotics for lasso-type estimators. *Annals of Statistics*, 28:1356–1378, 2000.
- [35] W. R. Mann. Mean value methods in iteration. *Proc. Amer. Math. Soc.*, 4:506–510, 1953.
- [36] N. Meinshausen. Relaxed lasso. *Computational Statistics and Data Analysis*, 52(1):374–393, 2007.

- [37] N. Meinshausen and P. Bühlmann. High dimensional graphs and variable selection with the lasso. *Annals of Statistics*, 34:1436–1462, 2006.
- [38] N. Meinshausen and B. Yu. Lasso-type recovery of sparse representations for high-dimensional data. Technical Report 720, Dept. of Statistics, UC Berkeley, 2006.
- [39] M. Nikolova. Local strong homogeneity of a regularized estimator. *SIAM Journal on Applied Mathematics*, 61(2):633–658, 2000.
- [40] A. Nolte and R. Schrader. A note on the finite time behavior of simulated annealing. *Math. Operat. Res.*, 25:476–484, 2000.
- [41] Z. Opial. Weak convergence of the sequence of successive approximations for nonexpansive mappings. *Bull. Amer. Math. Soc.*, 73:591–597, 1967.
- [42] M.R. Osborne, B. Presnell, and B.A. Turlach. On the LASSO and its dual. *J. Comput. Graph. Statist.*, 9(2):319–337, 2000.
- [43] A. B. Owen. A robust hybrid of lasso and ridge regression. *Prediction and Discovery (Contemporary Mathematics)*, 443:59–71, 2007.
- [44] P. Radchenko and G. James. Variable inclusion and shrinkage algorithms. *Journal of the American Statistical Association*. To appear.
- [45] Y. She. Improving lasso: Data-augmentation and weights. Technical report, Statistics Department, Stanford University, May 2007.
- [46] Y. She. Sparse regression with exact clustering. Technical report, Statistics Department, Stanford University, October 2007.
- [47] Y. She. Thresholding-based iterative selection procedures for model selection and shrinkage. Technical report, Statistics Department, Stanford University, June 2008.
- [48] K. Shimizu, Y. Ishizuka, and J.F. Bard. *Nondifferentiable and Two-Level Mathematical Programming*. Kluwer Academic Publishers, 1997.
- [49] J.-L. Starck, E. Candes, and D. Donoho. Astronomical image representation by the curvelet transform. *Astronomy and Astrophysics*, 398:785–800, 2003.

- [50] R. Tibshirani. Regression shrinkage and selection via the lasso. *JRSSB*, 58:267–288, 1996.
- [51] R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight. Sparsity and smoothness via the fused lasso. *JRSSB*, 67(1):91–108, 2005.
- [52] Z. Šidák. Rectangular confidence regions for the means of multivariate normal distribution. *JASA*, 62:626–633, 1967.
- [53] M. Wainwright. Sharp threshold for high dimensional and noisy recovery of sparsity. Technical report, Department of Statistics, University of California, Berkeley, 2006.
- [54] G. Wrinkler. An ergodic  $L^2$ -theorem for simulated annealing in Bayesian image reconstruction. *Journal of Applied Probability*, 28:779–791, 1990.
- [55] T. Wu and K. Lange. Coordinate descent algorithm for lasso penalized regression. *Ann. Appl. Stat.*, 2(1):224–244, 2008.
- [56] M Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *JRSSB*, 68:49–67, 2006.
- [57] M. Yuan and Y. Lin. On the non-negative garrotte estimator. *JRSSB*, 69, 2007. 143–161.
- [58] C.-H. Zhang and J. Huang. The sparsity and bias of the Lasso selection in high-dimensional linear regression. *Ann. Statist*, 36:1567–1594, 2008.
- [59] P. Zhao and B. Yu. Grouped and hierarchical model selection through composite absolute penalties. Technical report, Dept. of Statistics, University of California Berkeley, 2006.
- [60] P. Zhao and B. Yu. On model selection consistency of lasso. *Journal of Machine Learning Research*, 7:2541–2563, 2006.
- [61] H. Zou. The adaptive lasso and its oracle properties. *JASA*, 101(476):1418–1429, 2006.
- [62] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *JRSSB*, 67(2):301–320, 2005.
- [63] H. Zou and R. Li. One-step sparse estimates in nonconcave penalized likelihood models. *Annals of Statistics*, 2008. To appear.