New Results in Detection, Estimation, and Model Selection

A Thesis
Presented to
The Academic Faculty

by

Xuelei (Sherry) Ni

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

School of Industrial and System Engineering
Georgia Institute of Technology
May 2006

New Results in Detection, Estimation, and Model Selection

Approved by:

Professor Xiaoming Huo, Advisor
School of Industrial and Systems Engineering
Georgia Institute of Technology

Professor C. F. Jeff Wu
School of Industrial and Systems Engineering
Georgia Institute of Technology

Professor Ming Yuan
School of Industrial and Systems Engineering
Georgia Institute of Technology

Professor Brani Vidakovic
Department of Biomedical Engineering &
School of Industrial and Systems Engineering
Georgia Institute of Technology

Professor Liang Peng
School of Mathematics
Georgia Institute of Technology

Date Approved : Nov. 01, 2005

TABLE OF CONTENTS

LIST OF TABLES . . . . **vii**

LIST OF FIGURES . . . . **viii**

SUMMARY . . . . **xiii**

PART I DETECTABILITY OF CONVEX SETS

I INTRODUCTION **1**

1.1 Motivations 1

1.2 Statistical Model 2

1.3 Research Contributions 4

1.4 Organization of part I 4

II NUMBER OF CONVEX SETS IN A DIGITAL IMAGE **5**

2.1 Detection Method – the First Approach 5

2.2 Main Result 7

2.3 Proof of the Theorem 8

2.4 Conclusion 14

III MINIMAX PROPORTION OF AN H(OR V)-PARALLELOGRAM EMBEDDED IN A CONVEX SET **15**

3.1 Multiscale Detection of Convex Sets – the Second Approach 15

3.2 The Minimum Proportion of the Maximally Embedded hv-parallelogram . 18

3.3 Proof of the Main Theorem 18

3.3.1 Different Cases 18

3.3.2 Discussion Regarding the Foregoing Cases 23

3.4 Discussion and Conclusion 46

PART II REGRESSION MODELS

IV INTRODUCTION **47**

4.1 Linear Model 47

4.2 Contributions 49

4.3 Organization of part II 50

V RSBN: REGRESSION WITH STOCHASTICALLY BOUNDED NOISES 51

5.1 Introduction 51

5.2 Formulation and Main Theoretical Result 52

5.3 Regression Achieving Asymptotic Minimaxity 54

5.3.1 Asymptotic Normality 55

5.3.2 Minimum Asymptotic Variance Estimation 57

5.3.3 Least Informative Distribution 57

5.3.4 Regression with Stochastically Bounded Noises (RSBN) 59

5.3.5 Parameters in RSBN 60

5.3.6 Fisher Information of the Least Informative Distribution 62

5.3.7 Asymptotic Variance of RSBN 63

5.3.8 Robustness 64

5.4 Numerical Algorithm: Proximal Point Method 64

5.4.1 General Idea 64

5.4.2 Partial Inverse 65

5.4.3 Solving Partial Inverse 66

5.4.4 Application to RSBN 67

5.4.5 Analysis 68

5.5 Other Implementation: SNOPT and SQP 69

5.6 Simulation 70

5.6.1 An Illustrative Example: Variable Star 70

5.6.2 Comparison with Ordinary Least Square Estimate and Huber's Estimate 74

5.7 Discussion 80

5.7.1 A General Regression Formulation 80

5.7.2 Our Choice of Objective Function vs. Huber's Estimate 81

5.7.3 Other Theoretical Results 82

5.7.4 Convexity of Fisher Information $I(f)$ 82

5.7.5 Local Minimaxity 83

5.8 Conclusion 83

VI ACHIEVING OPTIMAL REPRESENTATION WITH STEPWISE ALGORITHMS IN REGRESSION . . . 84

6.1 Introduction . . . 85

6.2 Review of Two Optimization Problems . . . 87

6.2.1 Subset Selection Criteria and **(P0)** . . . 88

6.2.2 Greedy Algorithms and **(P1)** . . . 90

6.3 Motivations: Case Studies . . . 93

6.3.1 An Extremal Example for the Least Angle Regressions . . . 93

6.3.2 Variable Selection with Orthogonal Model Matrix . . . 97

6.4 Main Results: Conditions of Equivalence . . . 98

6.4.1 Sufficient and Necessary Conditions . . . 99

6.4.2 A Sufficient Condition for Covariates that are Mostly Correlated with the Response . . . 100

6.4.3 Sufficient Conditions based on the Model Matrix and the Correlations with Residuals . . . 103

6.5 Discussion . . . 110

6.5.1 Computing Versus Statistical Properties . . . 110

6.5.2 Other Works in Variable Selection . . . 110

6.5.3 Back Elimination . . . 111

6.5.4 Other Greedy Algorithms and Absolutely Optimal Subset in Variable Selection . . . 111

6.5.5 Other Related Topics . . . 112

6.6 Conclusion . . . 112

VII REGRESSIONS BY ENHANCED LEAPS-AND-BOUNDS VIA ADDITIONAL OPTIMALITY TESTS (LBOT) . . . 113

7.1 Introduction . . . 114

7.2 Review of Basics . . . 115

7.2.1 Computing Regarding Submodels . . . 116

7.2.2 Two Basic Linear Algebra Results . . . 116

7.3 Subset Arrangement and the Leaps-and-Bounds Algorithm . . . 117

7.3.1 Inverse Tree . . . 118

7.3.2 Pair Tree . . . 119

7.3.3 Test in the Original Leaps-and-Bounds Algorithm 121
7.4 Additional Optimality Tests 124
7.4.1 New Tests 124
7.4.2 Proof of Theorem 4.1 125
7.5 Algorithm 128
7.5.1 A Scheme to Scan Through the Pair Tree $PT(m)$ 128
7.5.2 Integrating the Optimality Tests 129
7.6 Simulations 129
7.6.1 Synthetic Data 129
7.6.2 Effect of Model 134
7.6.3 Heuristic: Pre-Sorting 134
7.6.4 Real Data 137
7.7 Discussion 137
7.8 Conclusion 139

APPENDIX A — PROOFS ASSOCIATED WITH CHAPTER VI (PART II) 140

APPENDIX B — PROOFS ASSOCIATED WITH CHAPTER VII (PART II) 143

REFERENCES 152

LIST OF TABLES

1 Some statistics for three regression methods for the distorted variable star data. 73

2 Pair tree and residual sums of squares. 130

3 Minimum residual sums of squares and the corresponding optimal k-subsets. 130

4 For different values of m, the average numbers of pairs that are examined by both LB and LBOT among ten random experiments are presented. The second column includes the total number of pairs in the complete pair trees. 131

5 The optimal subsets and corresponding residual sums of squares, for the diabetes data. 138

6 The optimal subsets and corresponding residual sums of squares for the Housing data. 138

LIST OF FIGURES

1 A convex set (a) and its embedding in a random field (b). 2

2 Notations for the second case. 8

3 Illustration for the proof of Lemma 3.1. 10

4 Illustration for the proof of Lemma 3.2. 10

5 Illustration of the proof of Lemma 3.3. 11

6 An example when minimax embedding 2/9 is achieved. 19

7 Illustration of a maximally embedded v-parallelogram 19

8 Illustration of the transformation, which transforms an arbitrary convex set into a convex set with a horizontal side. 20

9 Possible cases while projecting to the x-axis. TC stands for Temporary Case. 21

10 Subcases of Case C1. 22

11 Subcases of case C2. 23

12 Case 1-1 cannot occur. 24

13 Case 2-1 cannot occur. 25

14 Case C1-3 & Case C1-4, where Case C1-3 can be covered by another case and Case C1-4 is impossible. The shaded areas are the original MERs. (a) is for C1-3 and (b) is for C1-4. 25

15 Case 1-2: an overview. 26

16 Subcase of Case 1-2, where P_{31} is on the boundary of Ω' and $P_{22} \in \Omega'$. (a) demonstrates that P_{33} cannot be above l_{23}; (b) demonstrates that P_{33} cannot be below l_{23} when l_{33} intersects with x-axis outside of interval $(a, a+4\delta)$; (c) demonstrates that P_{33} cannot be below l_{23} when l_{33} intersects with x-axis inside of interval $(a, a+4\delta)$. 27

17 Subcase of Case 1-2, where P_{31} is on the boundary of Ω' and $P_{23} \in \Omega'$. (a) demonstrates that the larger the slope of l_{31} is, the larger the possible Ω' is; (b) demonstrates that P_{33} cannot be below or above l_{24} in order to make a larger feasible Ω'. 28

18 Subcase of Case 1-2, where P_{31} is on the boundary of Ω' and $P_{24} \in \Omega'$. (a) demonstrates that P_{23} should not be below or above l_{31} in order to have a larger Ω'; (b) demonstrates that P_{33} should not be below or above l_{25} in order to have a larger Ω'. 29

19 Subcase of Case 1-2, where P_{31} is on the boundary of Ω' and $P_{25} \in \Omega'$. (a) demonstrates that P_{24} should not be below or above l_{31} in order to have a larger and feasible Ω'; (b) demonstrates that the larger the absolute value of the slope of l_{33} is, the larger the possible Ω' is. 29

20 Subcase of Case 1-2, where P_{31} is on the boundary of Ω' and $P_{26} \in \Omega'$. This case is impossible because we can find a higher dyadic rectangle in Ω' with support $(a + 2\delta, a + 3\delta)$. 30

21 Subcase of Case 1-2, where P_{31} is on the boundary of Ω' and none of $P_{2j} \in \Omega'$ 30

22 Subcase of Case 1-2, where P_{32} is on the boundary of Ω'. In this case, the maximal Ω' is bounded by the x-axis, l_{32} which also passes through P_{46}, and the vertical line passes through point $(a, 0)$. 31

23 Subcase of Case 1-5, where P_{24} is on the boundary of Ω'. For the slope of l_{31}, Ω' is on the right of l_{31}, and the possible Ω' is larger when the slope of l_{31} is larger. 32

24 Subcase of Case 1-5, where P_{24} is on the boundary of Ω'. For the slope of l_{24}, comparing with the case where P_{33} is on l_{24}, (a) demonstrates that the maximal possible Ω' is smaller when P_{33} is above l_{24}; (b) demonstrates that the maximal possible Ω' is smaller when P_{33} is below l_{24}. 32

25 Subcase of Case 1-5, where P_{23} is on the boundary of Ω'. We are considering the slope of l_{25}. Comparing with the case where P_{33} is on l_{25}, (a) demonstrates that the maximal possible Ω' is smaller when P_{33} is above l_{25}; (b) demonstrates that the maximal possible Ω' is smaller when P_{33} is below l_{25}. 33

26 Subcase of Case 1-5, where P_{23} is on the boundary. We are considering the slope of l_{23}. Comparing with the case where P_{31} is on l_{23}, (a) demonstrates that the maximal possible Ω' is smaller when P_{31} is above l_{23}; (b) demonstrates that the maximal possible Ω' is smaller when P_{31} is below l_{23}. . . . 34

27 Case 2-2: an overview. Either P_{41} or P_{42} is on the boundary of Ω'. We assume it is P_{41}. For points P_{2j}, at most one of them will be inside of Ω'. . 34

28 Subcase of Case 2-2, where $P_{24} \in \Omega'$. This case cannot happen since a larger embedded dyadic rectangle can be found. 35

29 Subcase of Case 2-2, where $P_{25} \in \Omega'$. The slope of l_{41} is considered. It is clear that the larger the slope is, the larger the possible Ω' is. 35

30 Subcase of Case 2-2, where $P_{25} \in \Omega'$. The slope of l_{26} is considered. Comparing with the case where P_{34} is on l_{26}, (a) demonstrates that the maximal possible Ω' is smaller when P_{34} is above l_{26}; (b) demonstrates that the maximal possible Ω' is smaller when P_{34} is below l_{26}. 36

31 Subscase of case 2-2, where $P_{26} \in \Omega'$. Ω' is surrounded by l_{25}, l_{27} and l_{41}. (a) demonstrates that P_{34} cannot be below or above l_{27} in order to enclose a larger feasible Ω'. (b) demonstrates that P_{41} cannot be below or above l_{25} in order to have a larger feasible Ω'. 36

32 Subcase of Case 2-2, where $P_{27} \in \Omega'$. The maximal Ω' is limited by vertical l_{41}, vertical l_{34}, the x-axis, and l_{26}. Slope of l_{26} won't change the area of the maximal Ω' as long as it passes through P_{41}, or l_{41} is vertical. 37

33 Subcase of Case C2-3, where P_{32} is on the boundary of Ω'. None of P_{2j} is inside of Ω'. Ω' is surrounded by the vertical line l_a, the x-axis, and l_{32} that also passes P_{42}. 38

34 Subcase of Case C2-3, where P_{31} is on the boundary of Ω'. Region Ω' is between l_{31} and l_{33}. Note firstly, the larger the slope of l_{31}, the larger the possible Ω'. Secondly, $(a+2\delta, 0)$ is below l_{33}. However, the larger the distance between them, the smaller the possible Ω'. 38

35 Subcase of Case C2-4, where P_{33} is on the boundary of Ω'. Region Ω' is between l_{31} and l_{33}. Note firstly, the larger the slope of l_{31}, the larger the possible Ω' is. Secondly, $(a + 2\delta, 0)$ is below l_{33}. However, the larger the distance between them, the smaller the possible Ω'. 39

36 Subcase of Case C2-4, where P_{32} is on the boundary and $P_{24} \in \Omega'$. Ω' is bounded by l_{32}, l_{25}, l_{42}, and the x-axis. Line l_{32} is vertical because both P_{24} and P_{32} are on the line. l_{25} being zero and l_{42} being vertical will enclose larger Ω' that is applicable. 40

37 Subcase of Case C2-4, where P_{32} is on the boundary of Ω' and $P_{25} \in \Omega'$. Ω' is bounded by l_{32}, l_{26}, and the x-axis. For l_{32}, the larger the slope is, the larger the possible enclosed Ω' is. For l_{26}, point $(a + 2\delta, 0)$ is below l_{26}. The larger the distance between them is, the larger the possible Ω' is. Meanwhile, l_{26} is below P_{42} since P_{42} not in Ω'. Hence, the best l_{26} passes P_{42}. 41

38 Subcase of Case C2-2, where P_{32} is on the boundary and $P_{26} \in \Omega'$. Ω' is bounded by l_{32} and l_{27}. (a) indicates that in order to include a larger possible Ω', l_{32} should pass P_{25}, or more correctly, slightly below P_{25}; (b) indicates that l_{27} should pass P_{42} to include larger area. 41

39 Subcase of Case 2-4, where P_{32} is on the boundary and $P_{27} \in \Omega'$. Region Ω' is bounded by l_{32} and l_{28}. For l_{32}, when P_{26} is on it, the embedded area is larger than the area when P_{26} is not on the line. Line l_{28} should be vertical because P_{34} and P_{42} are not in Ω'. 42

40 Subcase of Case C2-4, where P_{32} is on the boundary of Ω' and none of $P_{2j} \in \Omega'$. Ω' is surrounded by l_{32}, l_{24}, and l_{28}. Line l_{28} is vertical because P_{34} is outside Ω' and $(a + 2\delta, 0)$ is inside. l_{24} is horizontal because l_{28} is vertical and more area are supposed to be enclosed. Hence, given l_{28} and l_{24}, the slope of l_{32} can be any positive value as long as $(a, 0)$ is above it. 42

41 Case C2-5 & Case C2-6. Two impossible cases because a larger embedded dyadic rectangle can be found with support $(a + 0.5\delta, a + \delta)$. 43

42 Case C2-7 & Case C2-8. (a) indicates that case C2-7 is covered by another case since comparing with the shaded part, a lower embedded rectangle with the same area but smaller height can be found; (b) indicates that case C2-8 is impossible because a larger dyadic rectangle can be found. 43

43 Subcase of Case C2-9, where P_{26} is on the boundary of Ω'. Region Ω' is between l_{32} and l_{26}. The slope of l_{32} is considered here. The line should be vertical such that the enclosed area is larger. 44

44 Subcase of Case 2-9, where P_{26} is on the boundary of Ω'. Region Ω' is between l_{32} and l_{26}. The slope of l_{26} is considered here. In order to include more area, (a) indicates that P_{34} cannot be above l_{26} and (b) indicates that P_{34} cannot be below l_{26}. 44

45 Subcase of Case 2-9, where P_{25} is on the boundary. Consider the slope of l_{27}, (a) indicates that the enclosed area is smaller if P_{34} is above l_{27}; (b) indicates that the enclosed area is smaller if P_{34} is below l_{27}. 45

46 Subcase of Case 2-9, where P_{25} is on the boundary. We consider the slope of l_{25}. (a) indicates that the enclosed area is smaller if P_{32} is above l_{25}; (b) indicates that the enclosed area is smaller if P_{32} is below l_{25}. 45

47 Objective function $\rho(x)$ in ordinary least square, Huber's M-estimate, and RSBN. 54

48 Parameter λ_1 vs. α. The upper one is ordinary; the bottom takes $\log 10$ on α. 62

49 Fisher information $I(f_0)$ versus α. The upper one takes ordinary coordinates; the lower takes $\log_{10}$ on α. 63

50 From top to bottom: (a) Integer-valued magnitude of a variable star at midnight on 600 successive nights; (b) The deviation (between estimation and observation) corresponding to the ordinary least square estimate; And (c) the deviation corresponding to RSBN. 71

51 The distorted variable star signal. On day 224 and 446, the decimal points in the observed values were wrongly shifted to the left (15 → 1.5) and right (16 → 160) respectively. Circles indicate the true values. 73

52 Differences between the original variable star signal and the estimates from distorted signal by three different methods. The corresponding methods are, from top to bottom: (a) ordinary least square regression, (b) RSBN, and (c) Huber's M-estimate with $\Delta = 0.5$. 74

53 (a) Standard Gaussian noise vector ε, (b) the deviation vector $d_{LS,1}$ of least square regression in the Gaussian noise ε case, (c) the distorted Gaussian noise vector ε', (d) the deviation vector $d_{LS,2}$ from the least square regression with the distorted Gaussian noise ε', (e) the corresponding deviation vector $d_{RSBN,2}$ from the RSBN, (f) the corresponding deviation vector $d_{Huber,2}$ from the Huber's estimate. 78

54 (a) The histogram of the ratios between the Huber's estimate and RSBN: $\|d^{(i)}_{Huber,2}\|_2^2/\|d^{(i)}_{RSBN,2}\|_2^2, i = 1, 2, \cdots, 1000$; (b) The histogram of the logarithm ratios $\log_{10}\left(\|d^{(i)}_{L_2,2}\|_2^2/\|d^{(i)}_{RSBN,2}\|_2^2\right), i = 1, 2, \cdots, 1000$, for the least square regression and the RSBN; (c) For Huber's estimate, number of deviations whose amplitudes are above the cut-off; (d) For the ordinary least square regression, the histogram of number of deviations whose amplitudes are above cut-off. 79

55 Graphical Illustration of Lasso Problem. 91

56 An inverse tree with $m = 5$. For simplicity, in figures, "1234" is equivalent with subset $\{1234\}$ in the text. 118

57 A pair tree with $m = 5$. 119

58 The number of pairs in LBOT versus the number of pairs used in LB. The number of covariates $m = 10$. 132

59 Histogram: when $m = 10$, the number of pairs in LBOT over the number of pairs used in LB. 133

60 Histograms: (a) and (c), the number of pairs in LB; (b) and (d), ratios between the numbers of pairs in LB and the numbers of pairs in LBOT. (a) and (b) (resp., (c) and (d)) are for Exp. A (resp., Exp. B). See the context for details. 135

61 Histograms: (a) and (c), the number of pairs in LB; (b) and (d), ratios between the numbers of pairs in LB and the numbers of pairs in LBOT. Subfigures (a) and (b) (resp., (c) and (d)) are for Exp. A (resp., Exp. C). See the context for details. 136

SUMMARY

This thesis contains two parts: the detectability of convex sets and the study on regression models.

In the first part of this dissertation, we investigate the problem of detecting the presence of an inhomogeneous region with a convex shape in a Gaussian random field. The first proposed detection method relies on checking a constructed statistic on each convex set within an $n \times n$ image. We prove that the number of convex sets grows faster than any finite-degree polynomial of n, which indicates that one approach of determining the asymptotic threshold of the detectability can not be adopted here. We then consider detecting hv-parallelograms instead of convex sets, which leads to a multiscale strategy that can have the order of complexity $O(n^2 \log(n))$. We prove that 2/9 is the minimum proportion of the maximally embedded hv-parallelogram in a convex set. Such a constant indicates the effectiveness of the above mentioned multiscale detection method in detecting convex sets.

In the second part, we study the robustness and the optimality of regression models, and propose an improved all-subset selection algorithm.

1. Firstly, for robustness, M-estimators in a regression model where the residuals are unknown but have stochastically bounded distribution are analyzed. An asymptotic minimax M-estimator is derived. The new method is named *regression with stochastically bounded noises* (RSBN). Simulations demonstrate the robustness of this approach, as well as advantages over commonly used estimates such as the ordinary least square estimate and the Huber's estimate. Insights from RSBN are discussed.

2. Secondly, for optimality, by analyzing the performance of *least angle regressions* (LARS) – a newly introduced stepwise algorithm for variable selection – we get interested in considering the conditions under which a vector is the solution of two

optimization problems. For these two problems, one can be solved by certain stepwise algorithms, the other is the objective function in many existing criteria in subset selection (including C_p, AIC, BIC, MDL, RIC, etc). The latter is proven to be NP-hard. Several conditions are derived. They give the conditions for a vector to be the common solution to the two optimization problems. When the conditions, which can be easily checked, are satisfied, a greedy algorithm can be used to solve the seemingly unsolvable problem.

3. Finally, extending the above idea to exhaustive subset selection in regression, we improve the widely-used algorithm – the leaps-and-bounds algorithm by Furnival and Wilson. The proposed method further reduces the number of subsets needed to be considered in the exhaustive subset search by considering not only the of residual sums of squares, but also the residuals, the model matrix, and the current coefficients.

PART I

Detectability of Convex Sets

CHAPTER I

INTRODUCTION

1.1 Motivations

Constantly improved imaging technology and cheaper and better computers give rise to demands of using digital images as tools for evaluation and analysis. Automatic analysis and extraction of information from an image becomes more and more important in many fields. In most of these applications, data (images) are collected by standard sensors, such as cameras and radars. Then, the collected images are analyzed for the detection and the recognition of the targets, either stationary or moving, with unknown background. Detecting an inhomogeneous region with a convex shape in a noisy environment is one of many problems.

We investigate detecting the presence of a convex set in a noisy digital image. Detecting such objects is not only a basic task for detecting more complex targets. It also plays an important role in a variety of areas, including medical imaging, satellite imaging, and so on. We list some of the applications in the following:

- In electron cryomicroscopy [100], accurate and automatic particle detection from cyro-electron microscopy (Cryo-EM) images is very important for fast and correct reconstruction of macromolecular structures. The goal of this step is to locate all particles, which always have elliptic or rectangular shapes, from the Cryo-EM images. Since achieving high-resolution reconstruction often requires over hundreds of thousands of particles, it is extremely important to design a fast and automatic algorithm.

- In geomorphology [57, 10, 97], impact crater detection and crater size frequency counting have a very high priority in Extra Terrestrial Mapping and planetary chronological research. For example, in the Mars exploration, the existence of numerous impact craters in one area will provide evidence on the evolving surface process on Mars, which may help us find the geological evidence for running water on, or just below, the surface of Mars, especially several billion years ago. Hence, automation of crater detection is an important initial step toward making more efficient the work of the analyst, who are facing huge volumes of images that are being obtained by missions.

- In medical science [56], accurately locate and isolate the lesions in a brain image or a skin image is crucial for accurate diagnosis. Detection of the lesions in the early stages will considerably reduces morbidity and mortality. However, automated detection is a challenging task due to several reasons: (a) low contrast between the lesion and the surrounding, (b) reflections and shadows due to wrong illumination, and (c) artifacts such as skin texture, air bubbles, and hair.

1.2 Statistical Model

We formulate the detection problem, and give a statistical model in this section. In order to illustrate the idea, Figure 1 (a) presents a convex set in a square and Figure 1 (b) illustrates the convex set in a noisy Gaussian random field. Suppose an image is sampled from a random field in this square, with the following property: inside the convex set, the Gaussian mean is slightly higher than the Gaussian mean outside. Question: how to detect the presence of such a convex set?

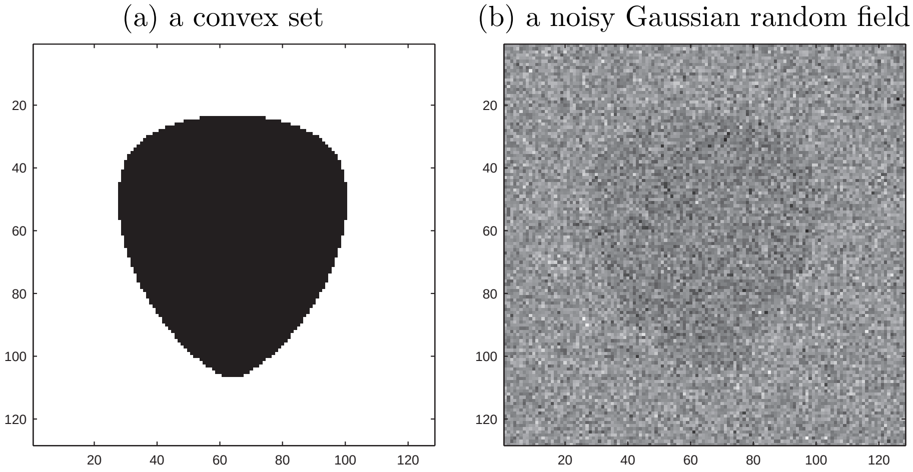


Figure 1: A convex set (a) and its embedding in a random field (b).

To formulate the problem statistically, we first introduce some notations for a digital image. An $n \times n$ digital image has double indices: (i, j), $0 \leq i, j \leq n - 1$. Each pair of indices indicate a pixel of the image. A subset of pixels is denoted by Ω, i.e., $\Omega \subset \{(i, j), 0 \leq i, j \leq n - 1\}$. A pixel p is called a *boundary pixel* of Ω, iff (if and only if) it belongs to Ω and one of its neighbor is outside of Ω. Ω is a convex set if and only if for any two points $x, y \in \Omega$, the line segment connecting x and y is inside the Ω. More rigorous definition for convex sets will be given later in Chapter 2 and Chapter 3, when we consider more specific detection approaches.

For a location p (with indices (i, j)), let $X(p)$ (or $X(i, j)$) denote the intensity of the image at pixel p, i.e., (i, j). We have

$$X(p) \sim \begin{cases} N(0, \sigma^2), & \text{if } p \notin \Omega, \\ N(\mu, \sigma^2), & \text{if } p \in \Omega, \end{cases}$$

where $N(\mu, \sigma^2)$ stands for a normal distribution with mean μ and standard deviation σ. An illustration of such a sampled image X is in Figure 1 (b). For future convenience, in this document, we assume $\sigma = 1$. That is, if the image has no embedded signal (i.e., a white noise image), then $X(i, j) \sim N(0, 1)$, for all $0 \leq i, j \leq n - 1$. Here $N(0, 1)$ stands for a normal distribution with mean 0 and variance 1. This situation is defined as the *null hypothesis* (denoted by $\mathbf{H_0}$). On the other hand, if there is a subset of pixels (denoted by Ω) satisfying that for a constant $\mu > 0$, $X(i, j) \sim N(\mu, 1)$ when pixel $(i, j) \in \Omega$, and $X(i, j) \sim N(0, 1)$ when pixel (i, j) is outside Ω, then Ω is an "embedded" object. Such a case is defined as an *alternative hypothesis* (denoted as $\mathbf{H_a}(\Omega, \mu)$). Note that by varying the subset Ω and the value of parameter μ, there are infinite number of possibilities for the alternative hypotheses. The objective of our detection problem is to decide whether or not such an object Ω exists. More specifically, how large should the value of μ and the area of Ω be so that the corresponding alternative hypothesis can be distinguished from the null hypothesis. In statistics, this is a typical hypothesis testing problem with a simple null hypothesis and a composite alternative hypothesis.

1.3 Research Contributions

The focus of this part of the dissertation is to give an efficient method to solve the hypothesis testing problem for the detection. In particular, we study and develop results that are able to derive meaningful and fast detection algorithms. The main contributions of this part can be summarized as follows:

- We propose one possible detection method based on the principle of the Generalized Likelihood Ratio Test (GLRT). The infeasibility of this approach is revealed by the study of the cardinality of the convex sets in an n by n image. More significantly, we give a recursive formula to compute the number of convex sets. From this formula, it can be proven that the number of convex sets grows faster than any finite-degree polynomial of n.

- We propose the second detection scheme based on a multiscale approach in detecting h(v)-parallelogram in an image. The efficiency of this procedure is analyzed by studying the minimax proportion of an h(or v)-parallelogram included in a convex set. We show that the proportion is a constant: 2/9. Hence, we provide a method that has the same testing power as detecting convex sets directly but having much lower order of complexity.

1.4 Organization of part I

The rest of part I is organized as follows.

- Chapter 2 proposes the first design of detection, counts the number of convex sets in a digital image, and shows the impracticality of this approach.

- Chapter 3 introduces the multiscale approach to detect rectangles or h(v)-parallelogram, proposes the second detection system, and shows the minimax proportion of a hv-parallelogram in a convex set.

CHAPTER II

NUMBER OF CONVEX SETS IN A DIGITAL IMAGE

In this chapter, we consider the number of convex sets in an n by n digital image. We prove that a finite degree polynomial solution does *not* exist. A recursive formula is provided. This problem is directly motivated by the signal detection problem in finding the inhomogeneous convex region in the image. However, due to the generality of the problem, it could have much wider impact.

This chapter is organized as follows. The detection scheme that motivated the research in this chapter is derived in Section 2.1. The main result is given in Section 2.2. The theorem is proved in Section 2.3. The report on our literature survey together with some concluding remarks are provided in Section 2.4.

2.1 Detection Method – the First Approach

In order to detect the existence of a significant area Ω in a noisy Gaussian random field, we consider the following hypothesis testing problem:

$H_0 : \quad X(i,j) \sim N(0,1)$ for all $0 \leq i,j \leq n-1$;

$H_a(\Omega,\mu) : \quad X(i,j) \sim N(\mu,1)$ for some $\mu > 0$ when $(i,j) \in \Omega$.

In this dissertation, we are interested in the case when Ω is a convex set. Recall the objective of the detection problem is to decide whether or not such an object Ω exists, so that the alternative hypothesis H_a can be distinguished from the null hypothesis H_0. The following is an approach that can be easily derived. A useful reference regarding this is [2]. The analysis is from an asymptotic viewpoint.

First, if Ω and μ is give, we have a simple null hypothesis versus a simple alternative. Define

$$X(\Omega) = \sum_{(i,j)\in\Omega} X(i,j)/\sqrt{|\Omega|},$$

where $|\Omega|$ is the number of pixels in Ω. Under H_0, it is not hard to derive that $X(\Omega) \sim$

$N(0,1)$, while under $H_a(\Omega,\mu)$, $X(\Omega) \sim N(\mu\sqrt{|\Omega|},1)$. Hence, one can easily conduct the likelihood ratio test of H_0 against $H_a(\Omega,\mu)$, simply by asking if

$$X(\Omega) > \tau,$$

for a threshold τ.

For the composite alternative hypothesis, where $\mu(>0)$ and Ω are both unknown, it's straightforward to consider the maximum among all $X(\Omega)$'s (denoted by X^*). I.e., we consider

$$X^* = \max_{\Omega \in \mathcal{F}_n} X(\Omega),$$

where $\mathcal{F}_n$ denotes a collection of all subsets that are under consideration. For example, when we consider the problem of detecting a convex set, the $\mathcal{F}_n = \{$all convex sets in an $n \times n$ image$\}$.

Now we derive a detection rule so that for the simple null and the composite alternative, the type-I error (i.e., Prob(reject $H_a|H_0$)) converges to 0 while the image size n goes to infinity. Given a constant $\tau > 0$, and taking advantage of a property of $N(0,1)$, we know

- under H_0, for any Ω, $P(X(\Omega) > \tau) < \frac{1}{\tau}e^{-\frac{1}{2}\tau^2}([80, page 191])$;

- moreover, $P(X^* > \tau) \leq |\mathcal{F}_n| \cdot P(X(\Omega) > \tau) \leq |\mathcal{F}_n|\frac{1}{\tau}e^{-\frac{1}{2}\tau^2}$. The first inequality is due to Bonferroni. The second one is a direct substitution. Here $|\mathcal{F}_n|$ is the cardinality of the set $\mathcal{F}_n$.

Notice that if $\tau^* = \sqrt{2\log|\mathcal{F}_n|} \to +\infty$, then under H_0, $P(X^* > \tau^*) \to 0$. This gives us a powerful hypothesis testing method. Namely, the probability of the type-I error of this test goes to zero. On the other hand, consider a subset Ω, within which there is a nonzero mean μ, we have $X(\Omega) \sim N(\mu\sqrt{|\Omega|},1)$. If the mean of this normal distribution $\mu\sqrt{|\Omega|} > \tau^*$ (respectively, $\mu\sqrt{|\Omega|} < \tau^*$), such a subset will (respectively, will *not*) be distinguishable from the null. Hence aforementioned choice of $\tau^* = \sqrt{2\log|\mathcal{F}_n|}$ gives a threshold on when a subset is *detectable.* Note the above argument implies an asymptotic argument: we skip the notion that $n \to \infty$.

Now we explain why a polynomial expression for the size of set $\mathcal{F}_n$ (i.e., $|\mathcal{F}_n|$) could be useful in determining the asymptotic detectability of convex sets. If the cardinality of set $\mathcal{F}_n$ can be a polynomial of image size n — for an integer $k > 0$, one has $|\mathcal{F}_n| = O(n^k)$, (i.e., $\lim_{n \to +\infty} \frac{|\mathcal{F}_n|}{n^k}$ = constant) — then $\tau^* = C_2\sqrt{2k \log n}$, where C_2 is a constant. Note that to increase the value of τ by a factor of 10, the value of n needs to be increased to n^{100}. The slow growth of τ, when $|\mathcal{F}_n|$ is a polynomial, is an interesting feature of this type of detection problems. The existence of a polynomial formula for the quantity $|\mathcal{F}_n|$ is of strong interest to us.

2.2 Main Result

We first establish a definition for convex sets. Note that due to the discreteness of the problem, there could be other ways to define a convex set.

Definition 2.1 (Convex Set) *A set Ω is convex iff (if and only if)*

1. *there exists a close chain of pixels: (a_1, b_1), (a_2, b_2), $\cdots$, (a_k, b_k), and (a_1, b_1), which belong to Ω, and their centers form the vertices of a convex non-degenerated polygon;*
2. *$\forall p \in \Omega$, the center of p is inside or on the boundary of the above mentioned polygon, and vice versa.*

We have clarify the importance of the cardinality of convex sets in an $n \times n$ digital image for evaluating the detectable threshold τ^*. We hope that the cardinality can be expressed in a polynomial of image size n. However, this is not the truth. In this chapter, a recursive formula is provided to compute the number of convex sets. Most significantly, the following theorem is proven.

.

Theorem 2.2 (Main) *Given the above definition, the number of convex sets increases faster than any finite degree polynomial of image size n, as $n \to \infty$.*

This result implies that the approach we introduced in Section 2.1 for determining the asymptotic threshold of the detectability of convex sets can not be adopted. However, we

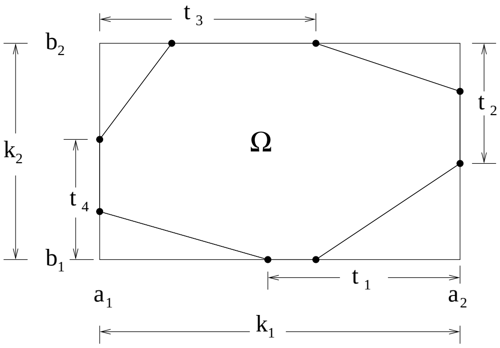


Figure 2: Notations for the second case.

would like to point out that even though Theorem 2.2 states that the number of convex sets is not polynomial, it would still be possible to have $\tau^* \sim \sqrt{\log(n)}$. In other words, the nonexistence of a polynomial formula merely invalidates a sufficient condition. The result of $\tau^* \sim \sqrt{\log(n)}$ can still be true. In fact, paper [2] gives a result of this kind. We refer to that paper for further details. Apparently, such a result can not be derived by counting the number of convex sets. Reference [2] also gives an excellent overview of the problem of detecting geometric objects in a random field.

2.3 Proof of the Theorem

We need some new notations. (Recall that a convex set is determined by a convex polygon whose vertices are the centers of some boundary pixels.) Let $a_1 = \min\{i : (i,j) \in \Omega\}$, $b_1 = \min\{j : (i,j) \in \Omega\}$, $b_2 = \max\{j : (i,j) \in \Omega\}$, and $a_2 = \max\{i : (i,j) \in \Omega\}$. The rectangle $[a_1, a_2] \times [b_1, b_2]$ is the minimum bounding rectangle of the convex set Ω. Let $t_1 = a_2 - \min\{i : (i, b_1) \in \Omega\}$, $t_2 = b_2 - \min\{j : (a_2, j) \in \Omega\}$, $t_3 = \max\{i : (i, b_2) \in \Omega\} - a_1$, and $t_4 = \max\{j : (a_1, j) \in \Omega\} - b_1$. An illustration is given in Figure 2.

We will need another notation: $H(a,b)$. For $a, b \geq 0$, a sequence of points — $(0,0)$, (c_1, d_1), (c_2, d_2), …, (c_ℓ, d_ℓ), (a,b) — determines a convex curve iff the chain of line segments, which connect these points by the same order, is convex. If this convex curve lies within the boundary of the right triangle with vertices $(0,0), (a,0)$, and (a,b) (boundary is included), we call it a *restricted convex curve* between $(0,0)$ and (a,b). Apparently, for a restricted convex curve, we must have $0 \leq c_1 \leq c_2 \leq \cdots \leq c_\ell \leq a$ and $0 \leq d_1 \leq d_2 \leq \cdots \leq d_\ell \leq b$.

More restrictively, if $\forall \ell, c_\ell < a$, we claim that this restricted convex curve does not intersect with the vertical line $i = a$. The total number of restricted convex curves that do not intersect with the vertical line $i = a$ is denoted by $H(a, b)$. Without much effort, one can derive

- $H(0, b) = 0$, for $b \geq 0$;
- $H(a, 0) = 1$, for $a \geq 1$; and
- $H(1, b) = 1$, for $b \geq 1$.

We would like to draw readers' attention to the fact that because $c_\ell < a$, the last segment ((c_ℓ, d_ℓ) to (a, b)) can *not* be the vertical line passing through point (a, b). Furthermore, readers may notice that under our definition, $H(a, b)$ and $H(b, a)$ could be unequal. For example, $H(0, b) \neq H(b, 0)$ when $b \geq 1$.

Recall Figure 2. It is not hard to prove that the following is the total number of convex sets under our definition:

$$\sum_{k_1, k_2 = 1}^{n} (n - k_1)(n - k_2) G(k_1, k_2), \tag{1}$$

where $k_1 = a_2 - a_1$, $k_2 = b_2 - b_1$, $G(k_1, k_2)$ is the number of convex sets whose minimal bounding rectangle is of size $k_1 \times k_2$. One can verify that, assuming $H(0, 0) = 1$,

$$G(k_1, k_2) = \sum_{\substack{0 \leq t_1, t_3 \leq k_1, \\ 0 \leq t_2, t_4 \leq k_2}} H(t_1, k_2 - t_2) H(t_2, k_1 - t_3) H(t_3, k_2 - t_4) H(t_4, k_1 - t_1). \tag{2}$$

Now the importance of $H(a, b)$ in our analysis is clear. To get our main result, we shall proceed by proving the following lemmas regarding $H(a, b)$.

Lemma 3.1 *The number of restricted convex curves between points $(0, 0)$ and (a, b), $a > b$ and with slopes < 1 is equal to $H(a - b, b)$. Here, "slopes" refer to the slopes of line segments that make up the convex curve.*

Proof. Readers can refer to Figure 3 for an illustration of the proof. First of all, the convex curves satisfying the condition of the lemma will lie within the triangle $((0, 0), (a -$

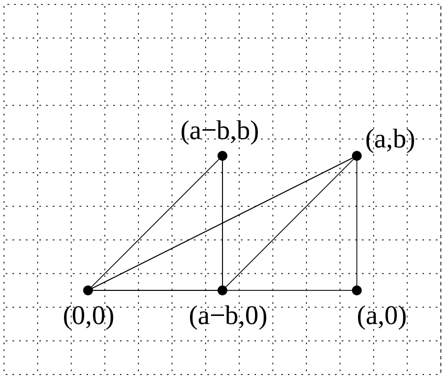


Figure 3: Illustration for the proof of Lemma 3.1.

$b, 0), (a, b))$, without touching the edge between $(a - b, 0)$ and (a, b), except the last point. For simplicity, we use C_1 to denote this set of convex curves . At the same time, $H(a - b, b)$ is the number of restricted convex curves between $(0, 0)$ and $(a - b, b)$ that do not intersect with line $i = a - b$. We use C_2 to denote this set of convex curves. We want to show $|C_1| = |C_2|$. Note that $\forall$ $\{(0, 0), (c_1, d_1), \ldots, (c_l, d_l), (a, b)\} \in C_1$, one can easily verify $\{(0, 0), (c_1 - d_1, d_1), \ldots, (c_l - d_l, d_l), (a - b, b)\} \in C_2$. On the other hand, $\forall$ $\{(0, 0), (e_1, f_1), \ldots, (e_m, f_m), (a-b, b)\} \in C_2$, $\{(0, 0), (e_1+f_1, f_1), \ldots, (e_m+f_m, f_m), (a, b)\} \in C_1$. Hence, there exists a one to one mapping between the curves in C_1 and the curves in C_2. The lemma is proved. □

Lemma 3.2 *The number of restricted convex curves that are between points* (0.0) *and* (a, b), $a < b$, *with slopes* ≥ 1, *and not intersecting with line* $i = a$, *is equal to* $H(a, b - a)$.

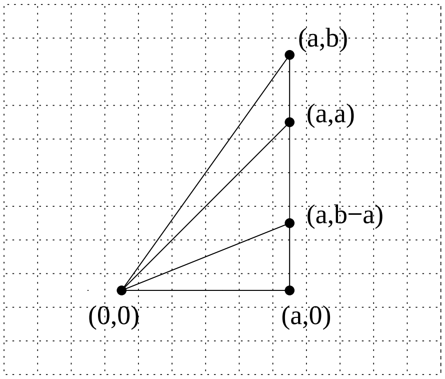


Figure 4: Illustration for the proof of Lemma 3.2.

Proof. This can be proved similarly with Lemma 3.1. We omitted the detail and only give the illustration in Figure 4.

For $H(a, b), b \geq a > 0$, we have the following recursive relation.

Lemma 3.3 (Recursive Rule) *For $b \geq a > 0$,*

$$H(a, b) = H(a, b - a) + \sum_{\substack{x_1 + x_2 \leq a, \\ x_1, x_2 \geq 1}} H(x_2, a - x_1 - x_2) H(x_1, b - a + x_2). \tag{3}$$

Proof. We describe it graphically. Refer to Figure 5.

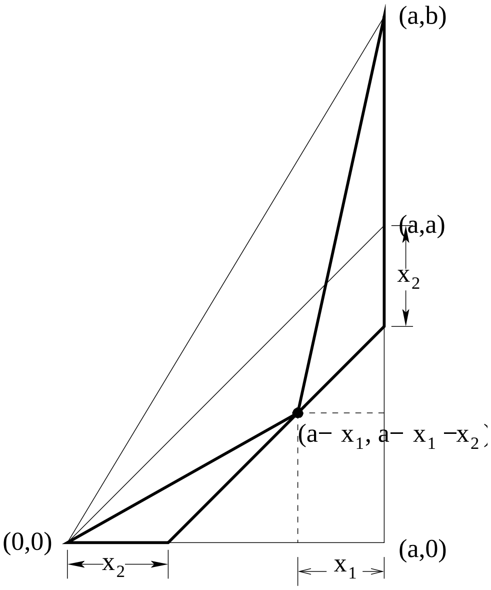


Figure 5: Illustration of the proof of Lemma 3.3.

For any curve that can be counted into $H(a, b)$, there are two possibilities (and only these two):

1. **Case 1.** The slopes of the curve are all ≥ 1.

2. **Case 2.** One of the vertices of the curves, (p^1, p^2), which is the center of a pixel p, satisfies the following: starting from the left, until reach its center, the slope of the convex curve is strictly less than 1; after this vertex, the slope of the convex curve is at least 1.

Hence,

$$\begin{aligned} H(a,b) &= \#\{\text{curves from Case 1}\} + \#\{\text{curves from Case 2}\} \\ &= \#\{\text{curves from Case 1}\} \\ &\quad + \sum_p \#\{\text{curves ending at } p\} \cdot \#\{\text{curves starting from } p\}. \end{aligned}$$

Under the first circumstance, the restricted convex curves have been analyzed in Lemma 3.2. So $\#\{\text{curves from Case 1}\} = H(a, b-a)$.

Under the second circumstance, since the slopes of the convex curve before (p^1, p^2) (including the edge ending at p) is strictly less than 1, (p^1, p^2) should be strictly under the line that connects $(0,0)$ and (a,a). I.e., $p^2 < p^1$. We can rewrite $p^2 = p^1 - x_2$, where $x_2 \geq 1$ and is illustrated in Figure 5. Also, since the convex curve cannot intersect with the vertical line $i = a$, p^1 should be strictly less than a. So we can rewrite $p^1 = a - x_1$, with integer $x_1 \geq 1$ that is also illustrated in Figure 5. The center of pixel p becomes $(a-x_1, a-x_1-x_2)$, $x_1 \geq 1, x_2 \geq 1$. At last, because $p^2 \geq 0$, we have $x_1 + x_2 \leq a$. Actually, one can check from Figure 5 that (p^1, p^2) can and only can lie strictly within the triangle with vertices $(0,0)$, (a,b), and $(a,0)$, or lie on the line segment connecting $(0,0)$ and $(a,0)$ (excluding the ending points). The geometric meanings of x_1 and x_2 are illustrated in Figure 5. Conditions $x_1 \geq 1, x_2 \geq 1$, and $x_1 + x_2 \leq a$ give an enumeration of all the possible positions of p.

Now, we have

$$H(a,b) = H(a,b-a) + \sum_{\substack{x_1+x_2 \leq a,\ x_1,x_2 \geq 1 \\ p=(a-x_1, a-x_1-x_2)}} \#\{\text{curves ending at } p\} \cdot \#\{\text{curves starting from } p\}.$$

From Lemma 3.1, the number of restricted convex curves between $(0,0)$ and $(a-x_1, a-x_1-x_2)$, with slopes < 1, is equal to $H((a-x_1)-(a-x_1-x_2), a-x_1-x_2) = H(x_2, a-x_1-x_2)$.

Now let's consider the last term, $\#\{\text{curves ending at } p\}$. By switching the origin $(0,0)$ to $(a-x_1, a-x_1-x_2)$, we observe that the number of restricted convex curves between $(a-x_1, a-x_1-x_2)$ and (a,b), with slopes ≥ 1, is equal to the number of convex curves between $(0,0)$ and $(a-(a-x_1), b-(a-x_1-x_2)) = (x_1, b-a+x_1+x_2)$, with slopes ≥ 1. The latter, from Lemma 3.2, is $H(x_1, b-a+x_2)$.

From all the above, the lemma is proved. □

As a direct application of Lemma 3.3, one can verify the following.

- $H(2,b) = 2 + \lfloor \frac{b-1}{2} \rfloor$, for $b \geq 1$ where $\lfloor x \rfloor$ is the largest integer that is no larger than x. This can be verified from $H(2,b) = H(2,b-2)+1$ for $b \geq 2$, which is stated by Lemma 3.3, and $H(2,0)=1$, $H(2,1)=2$.
- $H(3,b) = H(3, b - 3\lfloor \frac{b}{3} \rfloor) + 2\lfloor \frac{b}{3} \rfloor + \sum_{i=1}^{\lfloor \frac{b}{3} \rfloor} (\lfloor \frac{b+4-i}{2} \rfloor - i)$.
- $H(3,1)=3$, $H(3,2)=4$, and $H(3,3)=5$.

Another way to utilize Lemma 3.3 is to derive the following.

Corollary 3.4 *For $a \geq 1$,*

$$H(a,a) \geq 1 + \frac{1}{2}a(a-1).$$

Proof.

$$H(a,a) = H(a,0) + \sum_{\substack{x_1+x_2 \leq a \\ x_1,x_2 \geq 1}} H(x_2,\cdot)H(x_1,\cdot) \geq 1 + \sum_{x_1=1}^{a-1} \sum_{x_2=1}^{a-x_1} 1 = 1 + \frac{1}{2}a(a-1).$$

□

Recall we had $H(1,1)=1$, $H(2,2)=2$, $H(3,3)=5$.

We found that it is extremely difficult to derive a close form for the number of convex sets under our definition. This may explain why such a result does not exist in the published literature. In fact, by using Lemma 3.3, we can prove that if such a formula exists, it increases faster than any finite-degree polynomial of n, as n goes to infinity.

Proof of Theorem 2.2. From Lemma 3.3, one can prove for $a > 0$,

$$H(a,b) \geq \frac{b^{a-1}}{a^{2a}}. \qquad (4)$$

By choosing a large enough and $b = 2a^2$, the right hand side of (4) increases faster than any polynomial with a prescribed degree. Verifying (4) via (3) is a simple exercise. We describe it briefly below.

Proof of (4).

- When $b < a$,

$$H(a,b) > 1 > \frac{b^{a-1}}{a^{2a}}.$$

- When $b \geq a > 0$, from Lemma 3.3 and induction,

$$\begin{aligned}
H(a,b) &= H(a,b-a) + \sum_{\substack{x_1+x_2 \leq a,\\ x_1,x_2 \geq 1}} H(x_2, a-x_1-x_2)H(x_1, b-a+x_2) \\
&\geq \frac{(b-a)^{a-1}}{a^{2a}} + \sum_{\substack{x_1+x_2 \leq a,\\ x_1,x_2 \geq 1}} \frac{(a-x_1-x_2)^{x_2-1}}{x_2^{2x_2}} \frac{(b-a+x_2)^{x_1-1}}{x_1^{2x_1}} \\
&\geq \frac{(b-a)^{a-1}}{a^{2a}} + \left[\frac{(b-a)^{a-2}}{(a-1)^{2(a-1)}} + \frac{(b-1)^{a-3}}{(a-2)^{2(a-2)}} + \cdots\right] \\
&> \frac{1}{a^{2a}}\left[(b-a)^{a-1} + \binom{a}{1}a(b-a)^{a-2} + \binom{a}{2}a^2(b-a)^{a-3} + \cdots\right] \\
&= \frac{b^{a-1}}{a^{2a}}.
\end{aligned}$$

□

From (1) and (2), it is not hard to see that the number of convex sets also grows faster than any finite-degree polynomial of n. The theorem is proved.

2.4 Conclusion

The number of convex sets is a very general geometric problem. To our surprise, we can not locate any paper that directly address the problems being studied. The only remotely related work that we can find is [69], as well as some papers that were cited therein. There is a difference in the objective: they considered an algorithm for counting, instead of the cardinality of a collection of all convex sets.

Our motivation of studying this problem is from a detection problem that was described in the Introduction. However, as shown in this chapter, the number of convex sets grows faster than any finite-degree polynomial of power n (Theorem 2.2). This indicates that the introduced approach for determining the dectability of convex sets is not appropriate. But, as we indicated before, the theorem only shows the invalidation of a sufficient condition. Efficient detection scheme is still possible and we will give more results in the next chapter.

CHAPTER III

MINIMAX PROPORTION OF AN H(OR V)-PARALLELOGRAM EMBEDDED IN A CONVEX SET

Detecting the presence of a convex set in a Gaussian random field is considered further in this chapter. A multiscale strategy that is described in [2] can have the order of complexity $O(n^2 \log^2(n))$ for detecting a h(or v)-parallelogram in an n by n noisy image. So, instead of detecting convex sets directly, we can detect the h(or v)-parallelogram that embedded in convex regions with a nonzero Gaussian mean. We prove that 2/9 is the minimax proportion of a h(or v)-parallelogram included in a convex set. Such a constant indicates the effectiveness of the above mentioned multiscale detection method.

This chapter is organized as follows. Section 3.1 reviews the multiscale approach for detecting h(or v)-parallelograms. Section 3.2 gives the main result of this chapter. Section 3.3 develops the proofs for the theorem. Section 3.4 provides some discussions.

3.1 Multiscale Detection of Convex Sets – the Second Approach

Recall in the last chapter, we examined the following detection scheme. We calculate

$$X^* = \max_{\Omega \in \mathcal{F}_n} X(\Omega), \tag{5}$$

where Ω is a convex set, $\mathcal{F}_n$ is the collection of all convex sets, $X(\Omega) = \sum_{(i,j)\in\Omega} X(i,j)/\sqrt{|\Omega|}$. We hope to find X^* by enumerating all the convex sets in an image. Unfortunately, there is no numerically efficient way to compute the statistics $X(\cdot)$ for all convex sets. Mainly because there are too many convex sets.

In this chapter, we detect a more basic shape, named h(or v)-parallelogram, as a surrogate of detecting convex sets. It is relatively easy to compute the $X(\cdot)$-statistics for the new geometric objects. By investigating the relationship between a h(v)-parallelogram and a convex set, we can build a method to find inhomogeneous convex region indirectly.

An h-parallelogram was introduced in [2]. We give the definition and some related information in the following.

Definition 1.1 (h(or v)- parallelogram) *An h-(resp. v-) parallelogram is a parallelogram having two sides horizontal (resp. vertical) and its horizontal (resp. vertical) projection to the y- (resp. x-) axis on a Cartesian plane is a dyadic interval.*

A dyadic interval is defined as the following. Without loss of generality, we can assume the size of the image, n, equals to 2^m for some integer m. And we transfer the index set of pixels from $\{0, 1, ..., n-1\}$ to $\{0, 1/2^m, 2/2^m, ..., 1-1/2^m\}$. Then, a dyadic interval is defined as follows.

Definition 1.2 (dyadic interval) *Interval (a, b) is a dyadic interval if and only if there exist two non-negative integers s and ℓ, $s \le m$ and $\ell < 2^s$, such that $a = \ell/2^s$ and $b = (\ell+1)/2^s$.*

Now we reformat the testing scheme as the follows. Recall that for the image intensity, $X(i,j)$,

$$X(i,j) \sim \begin{cases} N(0,1), & \text{if } x \notin \Omega, \\ N(\mu,1), & \text{if } x \in \Omega, \end{cases}$$

where $\mu > 0$ and Ω is a convex set. Given a region $\widetilde{\Omega}$, we can calculate the statistic

$$X(\widetilde{\Omega}) = \sum_{(i,j)\in\widetilde{\Omega}} X(i,j)/\sqrt{|\widetilde{\Omega}|},$$

where $|\widetilde{\Omega}|$ denotes the number of pixels inside the set $\widetilde{\Omega}$. If set $\widetilde{\Omega}$ does not intersect with the “high activity” convex set Ω (i.e., $\widetilde{\Omega} \cap \Omega = \emptyset$), we have $X(\widetilde{\Omega}) \sim N(0,1)$. Otherwise, we have

$$X(\widetilde{\Omega}) \sim N\left(\mu \cdot \frac{|\widetilde{\Omega} \cap \Omega|}{\sqrt{|\widetilde{\Omega}|}}, 1\right),$$

where similarly the $|\widetilde{\Omega} \cap \Omega|$ denotes the cardinality of the intersection $\widetilde{\Omega} \cap \Omega$.

In Chapter 2, we choose the detection region $\widetilde{\Omega}$ as a convex set. In this chapter, we focus on h(v)-parallelograms. That is, we calculate

$$\widetilde{X}^* = \max_{\widetilde{\Omega} \text{ is an h(v)- parallelogram}} X(\widetilde{\Omega}). \tag{6}$$

It is provable that the above statistic is upper bounded by a quantity, which is a function of n. Specifically, if the null and alternative hypotheses are the following:

H_0 : $X(i, j) \sim N(0, 1)$ for all $0 \leq i, j \leq n - 1$;

$H_a(\widetilde{\Omega}, \mu)$: $X(i, j) \sim N(\mu, 1)$ for some $\mu > 0$ when $(i, j) \in \widetilde{\Omega}$.

It can be shown that there exists a constant Γ_n,

$$\frac{\Gamma_n}{\sqrt{2 \log(n^2)}} \to 1.$$

As $n \to \infty$ and the null hypothesis is true,

$$P\{\widetilde{X}^* < \Gamma_n\} \to 1.$$

That is, if we observe a $X(\widetilde{\Omega})$ that is larger than Γ_n, then the presence of an embedded h(v)-parallelogram can be claimed.

At resolution n, i.e., an n by n image, there are $O(n)$ dyadic intervals, including both vertical and horizontal directions. For each dyadic interval, there are at most $O(n^3)$ h(or v)-parallelograms: there are $O(n)$ options for two lower corners, at different side of a dyadic interval, the height of the parallelogram adds another $O(n)$ possibility. Hence the total number of h- (or v-) parallelograms is $O(n^4)$. Note that it is lower than the cardinality of all the convex set.

Hence, within $O(n^4)$ operations, we can detect the significant h(v)-parallelogram in an $n \times n$ image. Actually, a lower order algorithm can be derived by using a multiscale methodology with the help of Beamlets and Beamlet algorithms. This method is actually the idea in [2], where Arias-Castro et al. find that there exists an algorithm with order of complexity $O(n^2 \log^2(n))$ in detecting a h(v)-parallelogram. We omit the details and only mention the results with an emphasis that detecting h(v)-parallelogram can be done fast.

Note that we are interested in detecting convex sets, not a simple parallelogram. We should ask whether the above detecting rule can be adopted for convex regions. Furthermore, if yes to the previous question, how to adopt? We give the answers in the following section.

3.2 The Minimum Proportion of the Maximally Embedded hv-parallelogram

In this chapter, we analyze the relationship between an h(v)-parallelogram and a convex set. The main result is stated as the follows.

Theorem 2.1 (main theorem) *For any convex set, there is an embedded h- or v- parallelogram, which occupies at least* $2/9$ *of the convex set. Moreover, the constant* $2/9$ *can not be increased. In other words, for any quantity that is greater than* $2/9$*, there is a convex set, within which there is no embedded h- or v- parallelogram that takes the given proportion of the area of the convex set.*

This theorem is proved in continuum. In the discrete case, when the resolution $n \to \infty$, the same quantity holds.

Recall that we consider all the h- and v- parallelograms and have the new statistic: $\widetilde{X}^*$ as in (5). Comparing with X^* in (6), based on the above theorem, we can conclude

$$\widetilde{X}^* \asymp \sqrt{\frac{2}{9}} X^*$$

Hence an equally powerful test can be based on the comparison between $\frac{3}{\sqrt{2}}\widetilde{X}^*$ and Γ_n, which is given earlier.

3.3 Proof of the Main Theorem

The main theorem is proved in this section. We should consider both h-parallelograms and v-parallelograms. However, due to the symmetry of convex sets, only one type of parallelograms need to be considered. If we consider v-parallelogram alone, the minimax proportion 2/9 can be reached. One such limit case is shown in Figure 6. Without loss of generality, v-parallelograms are considered in the sequel.

3.3.1 Different Cases

A maximally embedded v-parallelogram is illustrated in Figure 7. Note that we do not use the word ‘inscribed’, to reflect the possibility that one corner point of a parallelogram may not be on the boundary of a convex set.

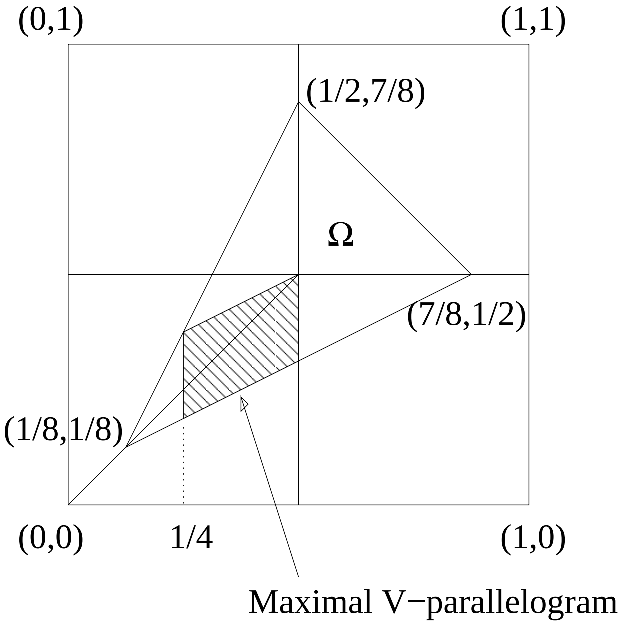


Figure 6: An example when minimax embedding 2/9 is achieved.

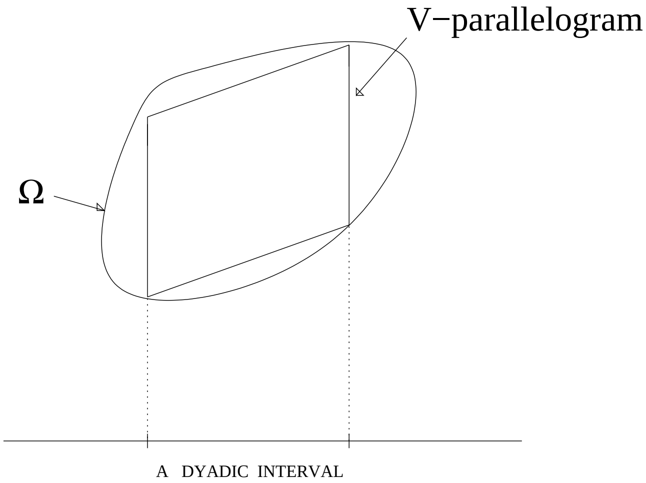


Figure 7: Illustration of a maximally embedded v-parallelogram

To simplify the proof, we assume that one side of the convex set is horizontal. An *affine transform* can be applied to convert an arbitrary convex set into a convex set with a horizontal side, as illustrated in Figure 8: the original set is Ω, and the transformed set is Ω'. Note the two sets have the same height at the same location. It is not hard to verify that Ω' is convex. Moreover, a maximally embedded v-parallelogram in Ω becomes a Maximally Embedded Rectangle (MER) in Ω'. Note that the rectangle must be supported by a dyadic interval on the x-axis, due to the definition of the v-parallelogram.

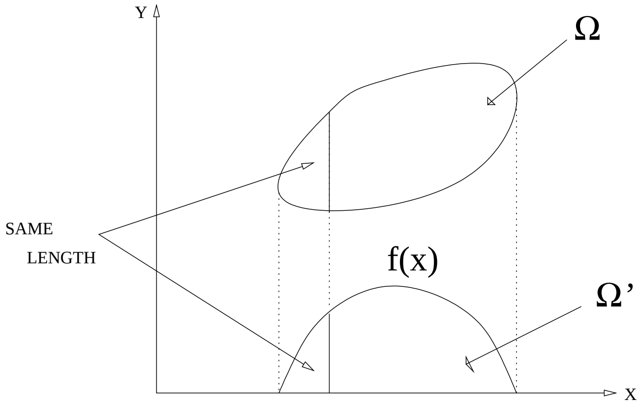


Figure 8: Illustration of the transformation, which transforms an arbitrary convex set into a convex set with a horizontal side.

The essence of the proof is to enumerate all the configurations of a convex set. We consider the horizontal side of a (transformed) convex set. Let a denote a dyadic number: i.e., there exist two integers s and ℓ, $\ell < 2^s$, such that $a = \ell/2^s$. Let $\delta = 1/2^{s'}, s' \geq s-1$. Note that intervals $(a, a+0.5\delta)$ and $(a+0.5\delta, a+\delta)$ are two dyadic intervals. For $(a, a+\delta)$, it is a dyadic interval when $s' \geq s$, and may not be when $s' = s-1$. We can always find an a and a δ such that (the horizontal side of) Ω' is complete inside of interval $(a, a+\delta)$, as shown in Figure 9 (a). We denote this case as TC-1. Now, if we consider the middle point $a+0.5\delta$, there are two possibilities:

1. If the middle point $a+0.5\delta$ is outside Ω', say, it is on the left of Ω', then by setting $a^{\text{new}} = a+0.5\delta$ and $\delta^{\text{new}} = \delta/2$, we go back to case TC-1 in Figure 9 (a). The case when $a+0.5\delta$ is on the right of Ω' can be similarly transferred to TC-1.

2. Therefore, we only need to consider the case when the middle point is inside Ω' (Figure

9 (c) TC-3).

Now we consider two more quarter points: $a + \delta/4$ and $a + 3\delta/4$. If none of them is inside Ω', which is illustrated in Figure 9 (d), let $a^{\text{new}} = a + \delta/4$ and $\delta^{\text{new}} = \delta/2$, we can transfer it back into case TC-3 in Figure 9 (c). So we only need to consider the case in which at least one of the above two points is inside Ω'. These two cases are illustrated in Figure 9 (e), in which both are inside, and Figure 9 (f), in which only one is inside. They are called cases C1 and C2, respectively, and will be investigated further.

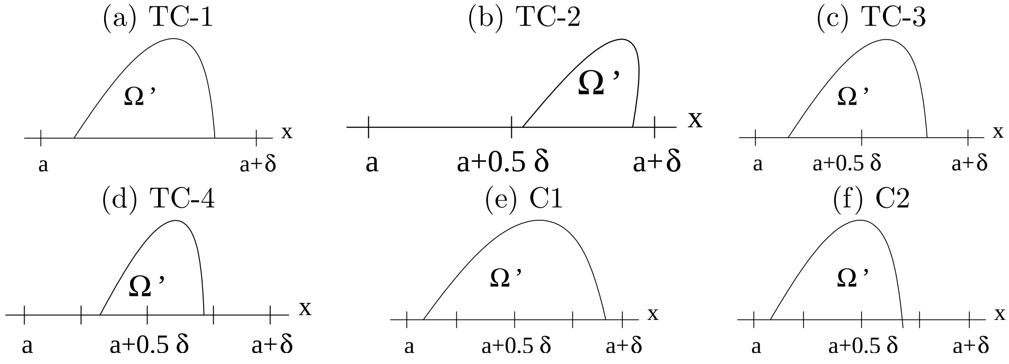


Figure 9: Possible cases while projecting to the x-axis. TC stands for Temporary Case.

We consider the MER. To reduce ambiguity, if there are two (embedded) rectangles with equal area, we always choose the one with larger support. Based on the definition of v-parallelogram, the support of MER (on the boundary of Ω' on x-axis) is a dyadic interval.

Within the case of C1, there are six subcases, as in the following table. The notations of points are illustrated in Figure 10. Note that there is a rescaling on the x-axis: $\delta^{\text{new}} = \delta/4$.

- C1-1: the support of the MER is with length $\leq \delta/4$, e.g., rectangle $P_{11}P_{12}P_{46}P_{45}$.

- C1-2: the support of the MER is with length δ, i.e., the support is $(a + \delta, a + 2\delta)$ or $(a + 2\delta, a + 3\delta)$. Due to symmetry, we only need to consider the MER with support $(a + \delta, a + 2\delta)$, which is rectangle $P_{31}P_{32}P_{44}P_{42}$ in the figure.

- C1-3: the support of the MER is with length $\delta/2$. This item and the next two cover this case. Due to symmetry, other locations are automatically taken care of. In this subcase, the MER is rectangle $P_{21}P_{22}P_{42}P_{41}$.

- C1-4: the support of the MER is with length $\delta/2$ and the MER is $P_{22}P_{23}P_{43}P_{42}$.
- C1-5: the support of the MER is with length $\delta/2$ and the MER is $P_{23}P_{24}P_{44}P_{43}$.

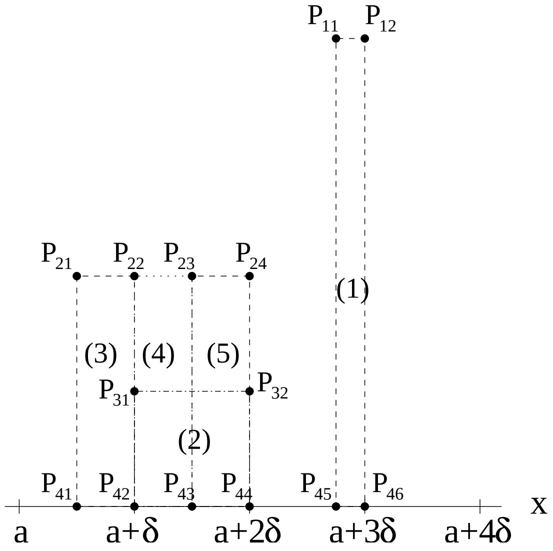


Figure 10: Subcases of Case C1.

Within C2, there are nine subcases, as illustrated in Figure 11.

C2-1: the support of the MER is with length $\leq \delta/8$, e.g. rectangle $P_{11}P_{12}P_{59}P_{58}$.

C2-2: the support of the MER is with length δ. The only possibility is $P_{41}P_{42}P_{57}P_{54}$.

C2-3: the support of the MER is with length $\delta/2$. Due to symmetry, only two conditions need to be considered. In this subcase, we consider rectangle $P_{31}P_{32}P_{54}P_{52}$.

C2-4: continuing from the above subcase, the MER is the rectangle $P_{32}P_{33}P_{56}P_{54}$.

C2-5: the support of the MER is with length $\delta/4$. Due to symmetry, five possibilities need to be considered. In this subcase, the MER is $P_{21}P_{22}P_{52}P_{51}$.

C2-6: the MER is $P_{22}P_{23}P_{53}P_{52}$.

C2-7: the MER is $P_{23}P_{24}P_{54}P_{53}$.

C2-8: the MER is $P_{24}P_{25}P_{55}P_{54}$.

C2-9: the MER is $P_{25}P_{26}P_{56}P_{55}$.

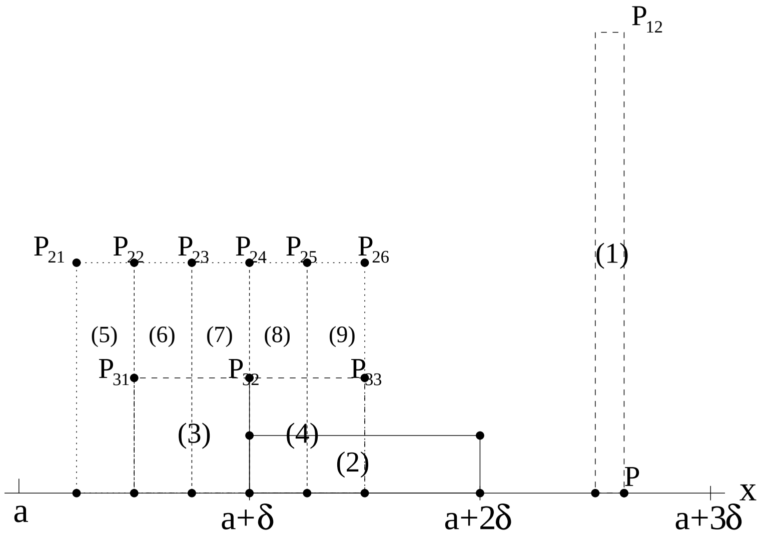


Figure 11: Subcases of case C2.

3.3.2 Discussion Regarding the Foregoing Cases

We prove the theorem through all the above subcases. All the proofs are illustrated in figures.

3.3.2.1 Case C1-1 and C2-1

For cases C1-1 and C2-1, it can be easily seen that these two subcases cannot exist. Actually, we can always find a contradiction. That is, inside the convex set, we can find other rectangles with dyadic supports and larger areas. These are illustrated in Figure 12 and 13, respectively.

To be more specific, we consider case C1-1 first. Under C1-1, the area of the MER candidate, rectangle $P_{11}P_{12}P_{46}P_{45}$, is less than or equal to $h\delta/4$ (shaded parts in Figure 12), where $\delta/4$ is the upper bound of the width and h is the height. The support can be either in interval $(a, a+\delta)$ or in $(a+\delta, a+2\delta)$, which also includes the other two possibilities $(a+2\delta, a+3\delta)$ and $(a+3\delta, a+4\delta)$ because of symmetry.

- When the support is within interval $(a+\delta, a+2\delta)$, this situation is illustrated in Figure 12 (a). From the definition of convex sets, it can be easily verified that the trapezoid with vertices $(a+\delta, 0)$, P_{11}, P_{12}, and $(a+3\delta, 0)$ is within the convex set Ω'. Hence, the rectangle embedded in the trapezoid with support $(a+\frac{3}{2}\delta, a+2\delta)$ and one

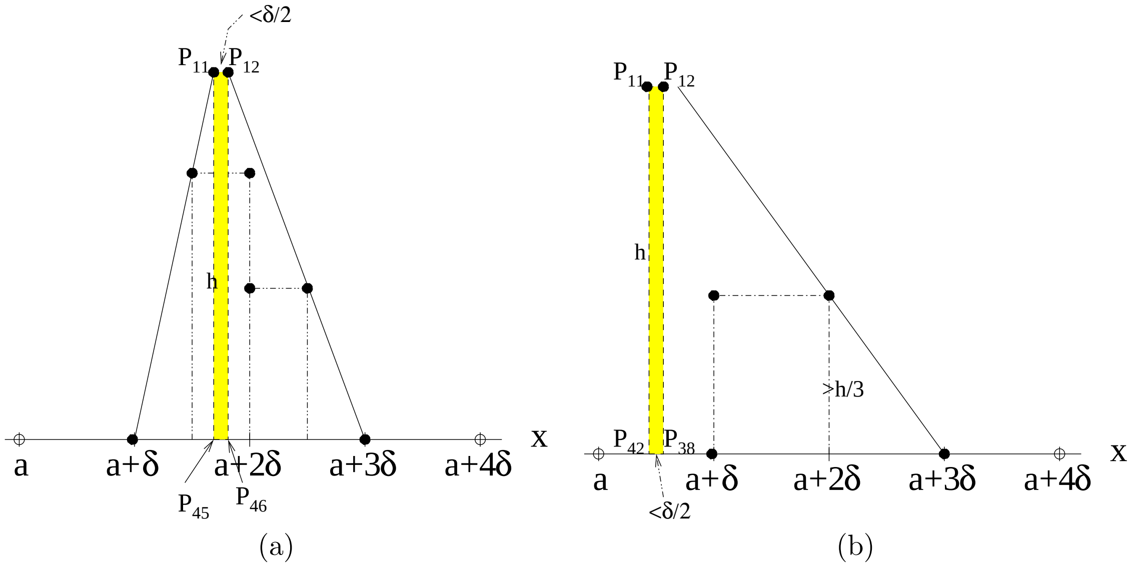


Figure 12: Case 1-1 cannot occur.

vertex on the line between $(a+\delta, 0)$ and P_{11} is inside Ω' as well. Note $(a+\frac{3}{2}\delta, a+2\delta)$ is a dyadic interval and the height of this new rectangle is greater than $h/2$, which leads to an area greater than $h\delta/4$. Hence, $P_{11}P_{12}P_{46}P_{45}$ cannot be the MER.

- When the support is within interval $(a, a+\delta)$, as illustrated in Figure 12 (b), an rectangle with dyadic support $(a+\delta, a+2\delta)$ can be found embedded inside Ω'. The height of this rectangle should be greater than $h/3$ by elementary knowledge in geometry. Hence, this embedded rectangle has area greater than $(h\delta/3)$, which also leads to a contradiction.

From all the above, case C1-1 does not exist.

Similarly, under the conditions of C2-1, no embedded rectangle with a dyadic base shorter than $\delta/8$ can be the MER, no matter where the rectangle is (cf. Figure 13).

3.3.2.2 Case C1-3 and C1-4

Cases C1-3 and C1-4 are similar with the above two cases C1-1 and C2-1. Under C1-3 (Figure 14 (a)), though, a shorter rectangle having the same area as the proposed MER can be found. The shorter rectangle also has a dyadic support and the original MER is shaded in the figure. Due to our preference for longer support, this one are embedded in another case. Under C1-4 (Figure 14 (b)), a larger embedded dyadic rectangle can be found.

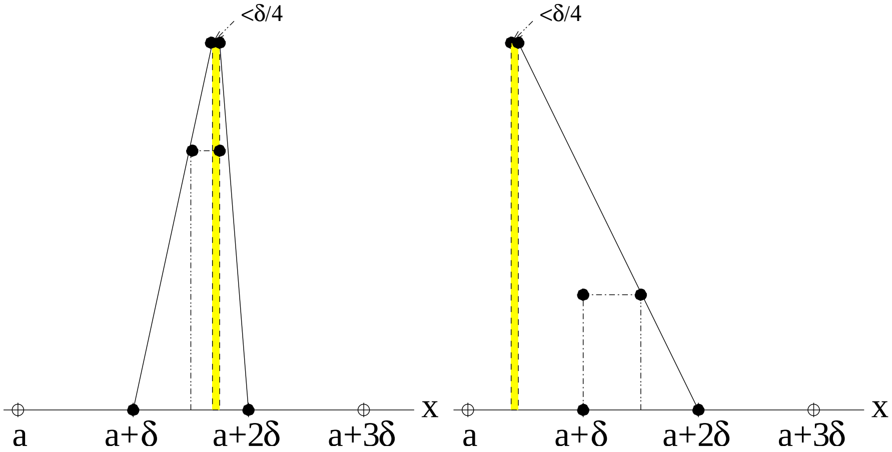
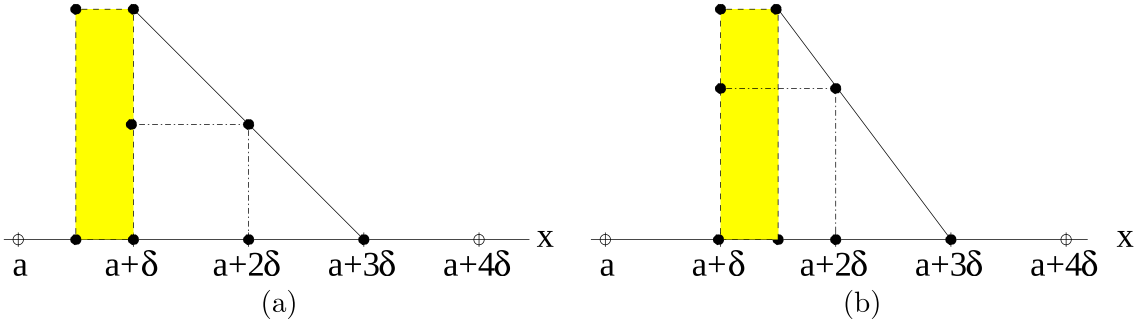


Figure 13: Case 2-1 cannot occur.



(a)

(b)

Figure 14: Case C1-3 & Case C1-4, where Case C1-3 can be covered by another case and Case C1-4 is impossible. The shaded areas are the original MERs. (a) is for C1-3 and (b) is for C1-4.

3.3.2.3 Case C1-2

Case C1-2 is much more complicated. We can divide this case further to get more detailed subcases. Figure 15 presents some key points that are important in the following discussion. In this figure and all the figures in the remaining of this report, a solid point means this point is inside or on the boundary of convex set Ω'. A circle either denotes a point outside of Ω', or a point we're not sure whether it is inside or outside. Notations P_{ij} are used to denote these points. Points $P_{i\times}$ are in the same height, while the height of points $P_{4\times}$ (given it is not on the x-axis) is half of the height of points $P_{3\times}$. Similarly, the height of points $P_{3\times}$ is half of the height of points $P_{2\times}$. For the horizontal inter-distance among $P_{\times j}$'s, if $x_{i,j}$ denotes the x-coordinate of P_{ij}, then intervals $(x_{i,j}, x_{i,j+1}), (x_{i,j+1}, x_{i,j+2})$ and so on are successive dyadic intervals with the same length. For the points at different level, the length of $(x_{i,j}, x_{i,j+1})$ is half of the length of interval $(x_{i+1,j}, x_{i+1,j+1})$. Moreover, l_{ij} denotes a line passing through point P_{ij} such that Ω' is on one side of the line.

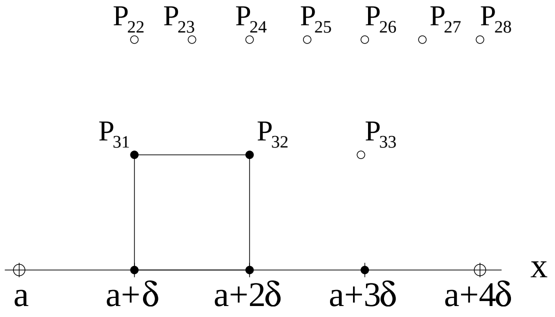


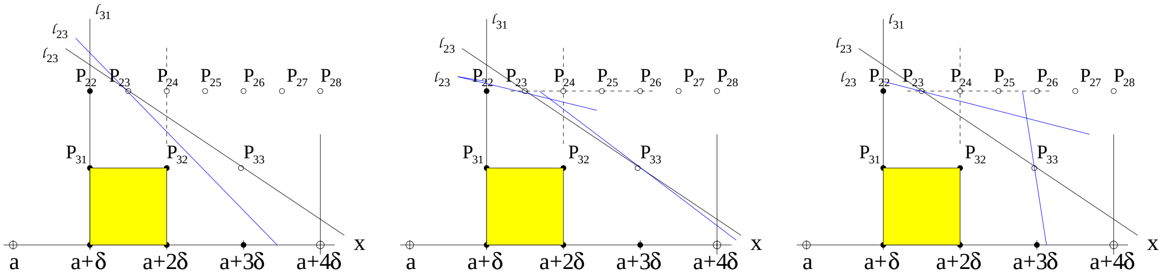
Figure 15: Case 1-2: an overview.

Now we return to case C1-2. Given Figure 15, where the MER have two vertices P_{31} and P_{32}, we know that at least one of the points P_{31} and P_{32} will be on the boundary of Ω'.

First, we assume P_{31} is on the boundary. Hence, line l_{31}, passing through P_{31}, can be chosen such that Ω' is on the right side of l_{31}. Moreover, P_{33} should be on the boundary of Ω' or $P_{33} \notin \Omega'$, i.e., it cannot be inside Ω. Otherwise, we can find a larger embedded rectangle with dyadic support $(a + 2\delta, a + 3\delta)$. Furthermore, among $P_{22}, P_{23}, P_{24}, P_{25}, P_{26}$ and P_{27}, at most one of them will be in the Ω'. Otherwise, we will have a contradiction regarding the MER again. Hence, we will have several subcases depending on the status of each P_{2j}.

If $P_{22} \in \Omega'$ (Figure 16), then $l_{31} \perp X$ and $P_{23} \notin \Omega'$. Hence, we can find a line l_{23} such that Ω' is on the left side of l_{23}. To make the possible Ω' have the maximal area, l_{23} should pass P_{33} as well. The reason is the following. Clearly, the Ω' with the maximal area is the triangle surrounded by l_{31}, l_{23}, and x-axis, or the quadrangle surrounded by those three lines and the additional side vertical to x-axis and parsing through $(a+4\delta-\varepsilon, 0)$. An offset ε is introduced because $(a+4\delta, 0) \notin \Omega'$. In Figure 16 (a), point P_{33} is either on l_{23} or above it. Obviously, the polygon with one side passing through P_{33} has larger area. In Figure 16 (b) and (c), P_{33} is either on l_{23} or below it. Difference between (b) and (c) is that in (b), line l_{33} intersects with x-axis outside interval $(a, a+4\delta)$; in (c), line l_{33} intersects with x-axis inside interval $(a, a+4\delta)$. Clearly, from the figures, both of (b) and (c) will give a larger Ω' when P_{33} is on l_{23}. Hence, in this case, the maximal Ω' is the quadrangle mentioned in this paragraph. We have, under this circumstance,

$$\frac{|MER|}{|\Omega'|} \geq \frac{1}{4} > \frac{2}{9}.$$



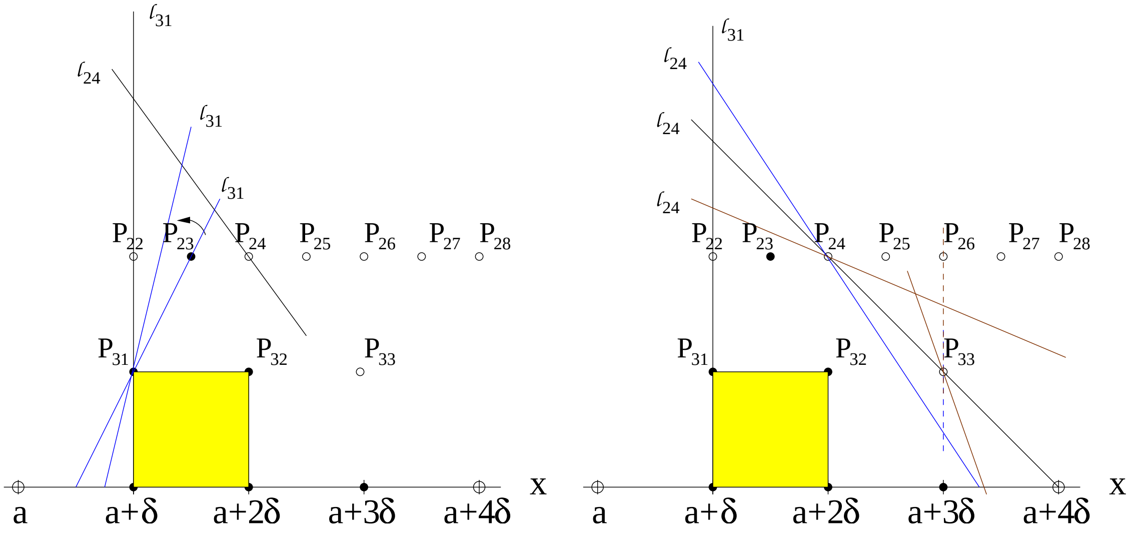
(a) P_{33} cannot be above l_{23} (b) P_{33} cannot be below l_{23} (c) P_{33} cannot be below l_{23}

Figure 16: Subcase of Case 1-2, where P_{31} is on the boundary of Ω' and $P_{22} \in \Omega'$. (a) demonstrates that P_{33} cannot be above l_{23}; (b) demonstrates that P_{33} cannot be below l_{23} when l_{33} intersects with x-axis outside of interval $(a, a+4\delta)$; (c) demonstrates that P_{33} cannot be below l_{23} when l_{33} intersects with x-axis inside of interval $(a, a+4\delta)$.

If $P_{23} \in \Omega'$ (Figure 17), then $P_{24} \notin \Omega'$. Recall P_{31} is on the boundary of Ω'. We have that Ω' is on the right side of l_{31} and the left side of l_{24}. Through Figure 17 (a), we find that when the slope of l_{31} is increasing, the area of the possible Ω' is increasing. Through

Figure 17 (b), we find that l_{24} should pass through P_{33}. Hence, Ω' is within the triangle bounded by the l_{31} that is vertical to the x-axis, the l_{24} that passes through P_{33} and the x-axis. Hence, in this case,

$$\frac{|MER|}{|\Omega'|} \geq \frac{2}{9}.$$



(a) slope of l_{31}

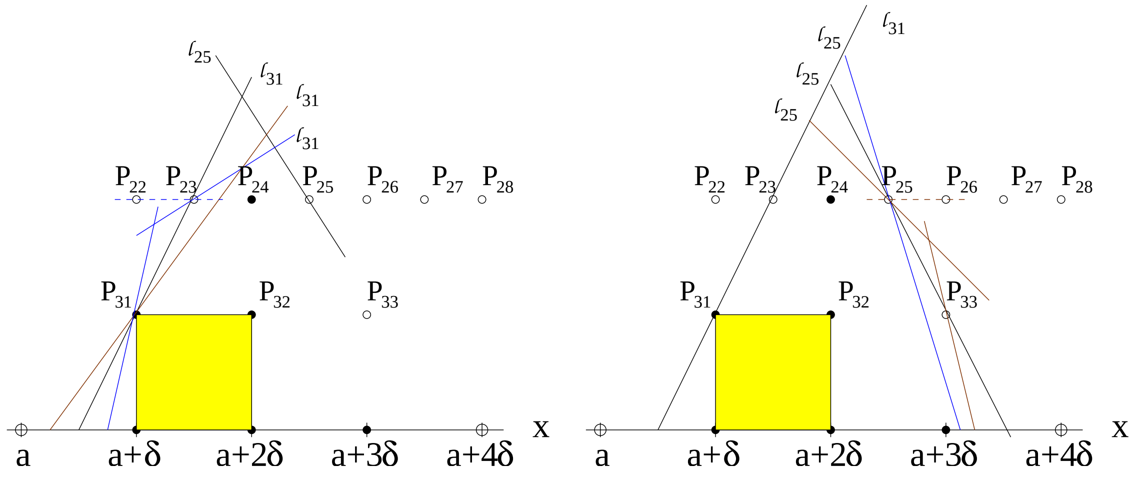
(b) slope of l_{24}

Figure 17: Subcase of Case 1-2, where P_{31} is on the boundary of Ω' and $P_{23} \in \Omega'$. (a) demonstrates that the larger the slope of l_{31} is, the larger the possible Ω' is; (b) demonstrates that P_{33} cannot be below or above l_{24} in order to make a larger feasible Ω'.

If $P_{24} \in \Omega'$ (Figure 18), we have $P_{25} \notin \Omega'$. Hence, Ω' is on the right side of l_{31} and the left side of l_{25}. The maximal and applicable Ω' should be within the triangle bounded by l_{31}, l_{25}, and the x-axis, where l_{31} should pass through P_{23} (Figure 18 (a)) and l_{25} should pass through P_{33} (Figure 18 (b)). Therefore,

$$\frac{|MER|}{|\Omega'|} \geq \frac{2}{9}.$$

If $P_{25} \in \Omega'$ (Figure 19), Ω' is on the right side of l_{31} and the left side of l_{33}. We study l_{33} instead of l_{26} because $P_{26} \notin \Omega'$, P_{33} is on the boundary of Ω' or $P_{33} \notin \Omega'$, and P_{33} is exactly below P_{26}. We observe that the possible Ω' is limited by l_{31}, which passes through P_{24} (Figure 19 (a)), l_{33}, which also passes through P_{26} (Figure 19 (b)), and the x-axis.

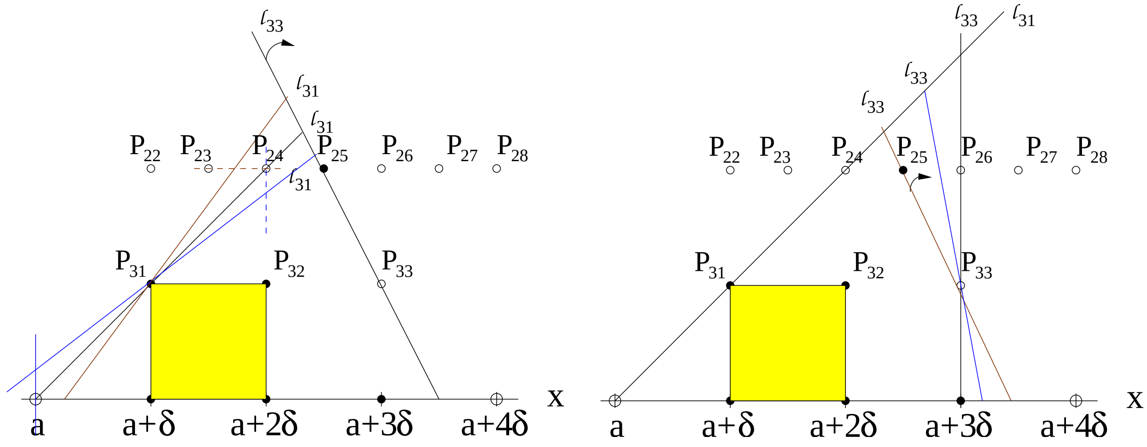


(a) slope of l_{31} (b) slope of l_{25}

Figure 18: Subcase of Case 1-2, where P_{31} is on the boundary of Ω' and $P_{24} \in \Omega'$. (a) demonstrates that P_{23} should not be below or above l_{31} in order to have a larger Ω'; (b) demonstrates that P_{33} should not be below or above l_{25} in order to have a larger Ω'.

Consequently, we have

$$\frac{|MER|}{|\Omega'|} \geq \frac{2}{9}.$$



(a) slope of l_{31} (b) slope of l_{33}

Figure 19: Subcase of Case 1-2, where P_{31} is on the boundary of Ω' and $P_{25} \in \Omega'$. (a) demonstrates that P_{24} should not be below or above l_{31} in order to have a larger and feasible Ω'; (b) demonstrates that the larger the absolute value of the slope of l_{33} is, the larger the possible Ω' is.

If $P_{26} \in \Omega'$ (Figure 20), we can prove that such a case is impossible by finding a larger dyadic rectangle in Ω' with support $(a+2\delta, a+3\delta)$. The same thing will happen if $P_{27} \in \Omega'$.

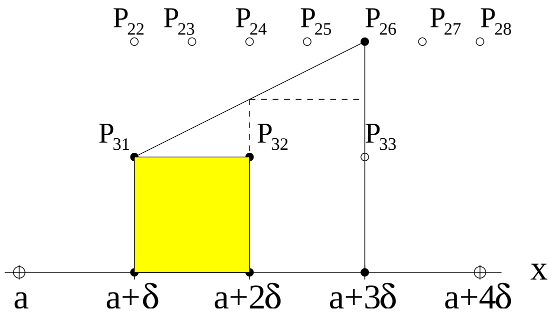


Figure 20: Subcase of Case 1-2, where P_{31} is on the boundary of Ω' and $P_{26} \in \Omega'$. This case is impossible because we can find a higher dyadic rectangle in Ω' with support $(a+2\delta, a+3\delta)$.

If none of $P_{2x} \in \Omega'$ (Figure 21). It is obvious that the maximal Ω' is smaller than the previous subcases. Possible Ω''s are illustrated in the figure. Hence,

$$\frac{|MER|}{|\Omega'|} \geq \frac{1}{4} > \frac{2}{9}.$$

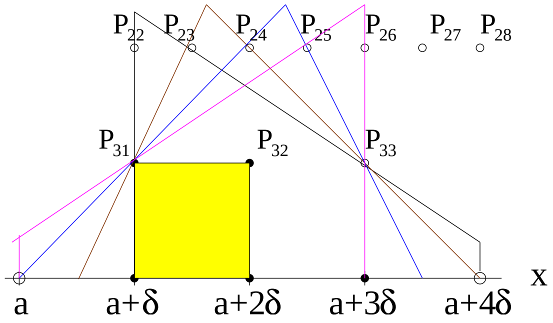


Figure 21: Subcase of Case 1-2, where P_{31} is on the boundary of Ω' and none of $P_{2j} \in \Omega'$

All the above are under the condition that P_{31} is on the boundary of Ω'. One the other hand, when P_{32} is on the boundary of Ω', it is much simpler. Reader can refer to Figure 22 for more details. More specifically, since $P_{46} \in \Omega'$ and P_{32} is on the boundary, none of $P_{2x}(x \geq 2)$ is in Ω' and the maximal possible Ω' is bounded by the x-axis, l_{32}, which also passes through P_{46}, and the vertical line passes through point $(a, 0)$. Hence, we have

$$\frac{|MER|}{|\Omega'|} \geq \frac{2}{9}.$$

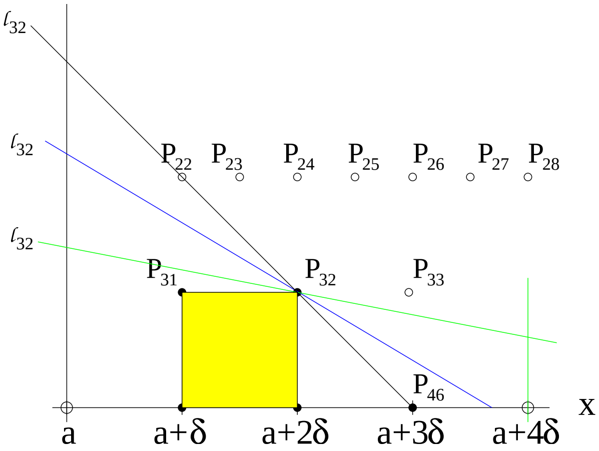


Figure 22: Subcase of Case 1-2, where P_{32} is on the boundary of Ω'. In this case, the maximal Ω' is bounded by the x-axis, l_{32} which also passes through P_{46}, and the vertical line passes through point $(a, 0)$.

We have finished the discussion about case C1-2. The analysis for the other cases is similar. We will just briefly go through the proof. Readers should be able to figure out the details by referring to the figures.

3.3.2.4 Case C1-5

For case C1-5, it can be subdivided as follows. Two points are critical: point P_{23} and point P_{24}. For these two points, at least one of them should be on the boundary of Ω'.

We first assume that point P_{24} is on the boundary of Ω' (Figure 23 and Figure 24). Hence, Ω' is on the left of certain line l_{24}, which passes through P_{24}. Furthermore, $P_{31} \notin \Omega'$, otherwise, the MER has support $(a+\delta, a+2\delta)$. So, Ω' is on the right of the line l_{31}. Figure 23 shows that the possible Ω' is larger when the slope of l_{31} is larger. Figure 24 details that l_{24} should pass though P_{33} (actually, a little below P_{33} since $P_{33} \notin \Omega'$) in order to get larger Ω'. Hence, the maximal possible Ω' is surrounded by vertical line l_{31}, the x-axis, and l_{24} which also passes through P_{33}. So, we have

$$\frac{|MER|}{|\Omega'|} \geq \frac{2}{9}.$$

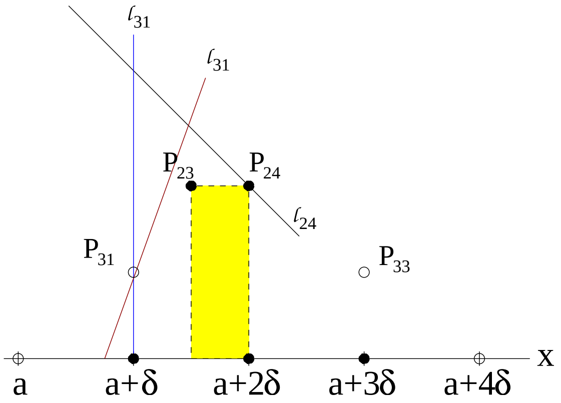
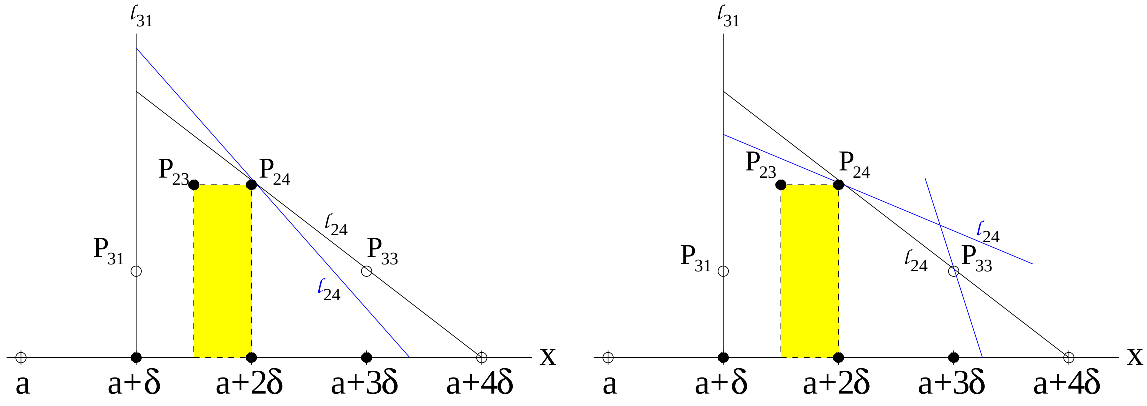


Figure 23: Subcase of Case 1-5, where P_{24} is on the boundary of Ω'. For the slope of l_{31}, Ω' is on the right of l_{31}, and the possible Ω' is larger when the slope of l_{31} is larger.



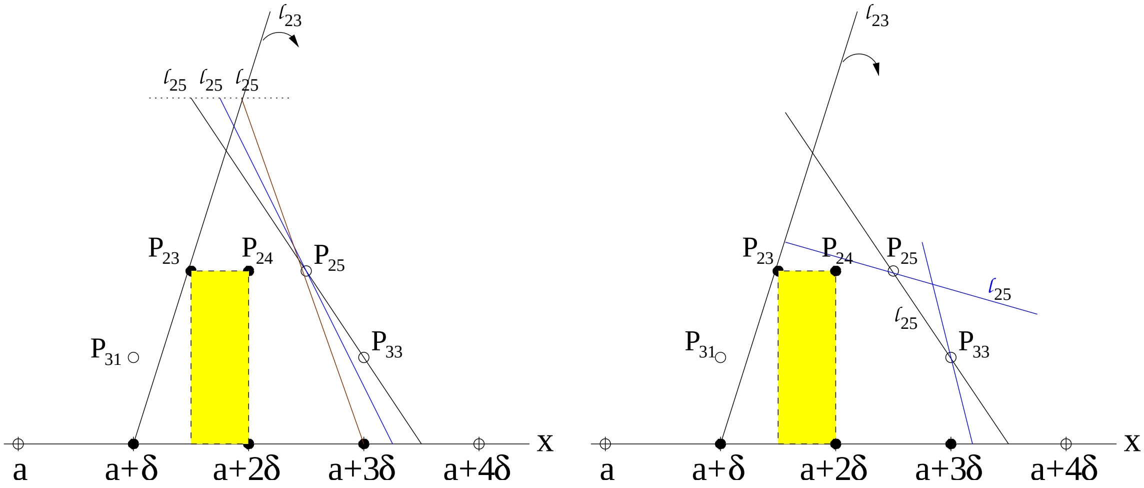
(a) P_{33} cannot be above l_{24}

(b) P_{33} cannot be below l_{24}

Figure 24: Subcase of Case 1-5, where P_{24} is on the boundary of Ω'. For the slope of l_{24}, comparing with the case where P_{33} is on l_{24}, (a) demonstrates that the maximal possible Ω' is smaller when P_{33} is above l_{24}; (b) demonstrates that the maximal possible Ω' is smaller when P_{33} is below l_{24}.

Next, if we assume point P_{23}, not point P_{24}, is on the boundary of Ω' (Figure 25 and Figure 26). Hence, Ω' is on the right of certain line l_{23}. Furthermore, P_{25} is not inside Ω', which means that there exists a line l_{25} such that Ω' is on the left of it. Note l_{23} and l_{25} are critical here. From Figure 25, we observe that l_{25}, making a larger Ω', should pass through P_{33}. From Figure 26, we observe that l_{23}, making a larger Ω', should also pass through P_{31}. Hence, the maximal possible Ω' is surrounded by these two specified lines and the x-axis. We have

$$\frac{|MER|}{|\Omega'|} \geq \frac{2}{9}.$$



(a) P_{33} cannot be above l_{25}

(b) P_{33} cannot be below l_{25}

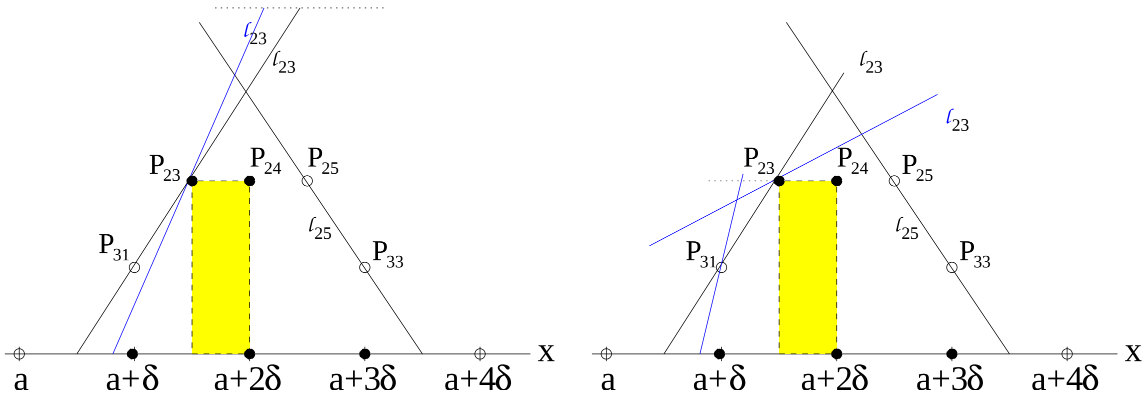
Figure 25: Subcase of Case 1-5, where P_{23} is on the boundary of Ω'. We are considering the slope of l_{25}. Comparing with the case where P_{33} is on l_{25}, (a) demonstrates that the maximal possible Ω' is smaller when P_{33} is above l_{25}; (b) demonstrates that the maximal possible Ω' is smaller when P_{33} is below l_{25}.

We finish case C1-5 and conclude that under this case, the theorem holds.

3.3.2.5 Case C2-2

Now we look at case C2-2, which is quite similar with case C1-2, but more complicated. The reason is we have more P_{2j}'s to be considered.

With an overview of this case (Figure 27), we know either P_{41} or P_{42} will be on the boundary of Ω'. Due to the symmetry, these two are the same and we assume that P_{41} is on the boundary. Furthermore, among P_{24} up to P_{211}, at most one of them will be inside



(a) P_{31} cannot be above l_{23} (b) P_{31} cannot be below l_{23}

Figure 26: Subcase of Case 1-5, where P_{23} is on the boundary. We are considering the slope of l_{23}. Comparing with the case where P_{31} is on l_{23}, (a) demonstrates that the maximal possible Ω' is smaller when P_{31} is above l_{23}; (b) demonstrates that the maximal possible Ω' is smaller when P_{31} is below l_{23}.

of Ω'. Hence, we will have several subcases with respect to the state of each P_{2j}. We deal with them in the following.

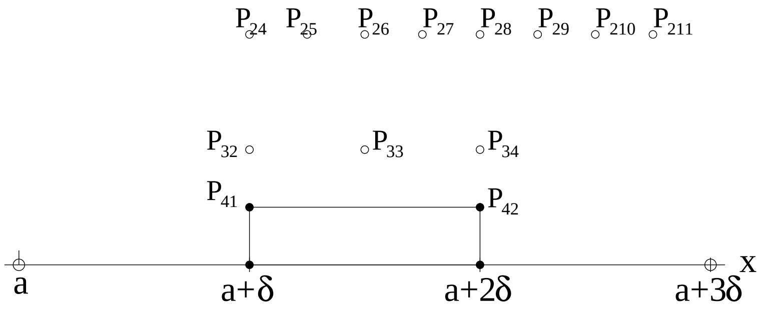


Figure 27: Case 2-2: an overview. Either P_{41} or P_{42} is on the boundary of Ω'. We assume it is P_{41}. For points P_{2j}, at most one of them will be inside of Ω'.

If $P_{24} \in \Omega'$ (Figure 28), we can prove this case is impossible since we can find a larger dyadic rectangle in Ω' with support $(a+\delta, a+0.5\delta)$. Similarly, the cases where $P_{28} \in \Omega'$, or $P_{29} \in \Omega'$, or $P_{210} \in \Omega'$, or $P_{211} \in \Omega'$ contradict with the assumption of the MER. Hence, only 4 subcases need to be analyzed.

If $P_{25} \in \Omega'$ (Figure 29 and Figure 30), then P_{26} is the closest point to P_{25} among the P_{2j}'s and $P_{26} \notin \Omega'$. Hence, in this case, the critical lines are l_{26} and l_{41}, where Ω' is on the left side of certain l_{26} and the right hand side of certain l_{41}. Figure 29 deals with the

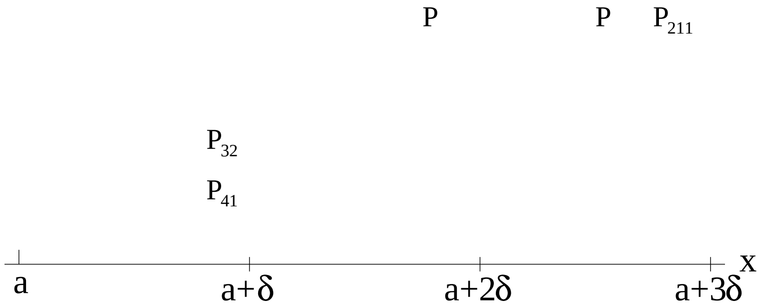


Figure 28: Subcase of Case 2-2, where $P_{24} \in \Omega'$. This case cannot happen since a larger embedded dyadic rectangle can be found.

slope of l_{41} and Figure 30 deals with the slope of l_{26}. From Figure 29, we observe that the larger the slope of l_{41}, the larger the possible Ω'. From Figure 30, we observe that l_{26} should pass through P_{34} to enclose a larger Ω'. Hence, the maximal possible Ω' is enclosed by the vertical l_{41}, the x-axis, and l_{26} that also passes through P_{34}. Hence, we have

$$\frac{|MER|}{|\Omega'|} \geq \frac{2}{9}.$$

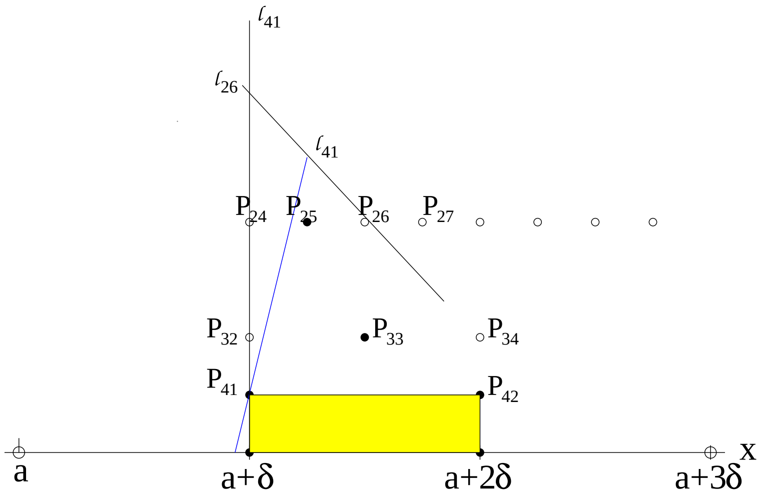
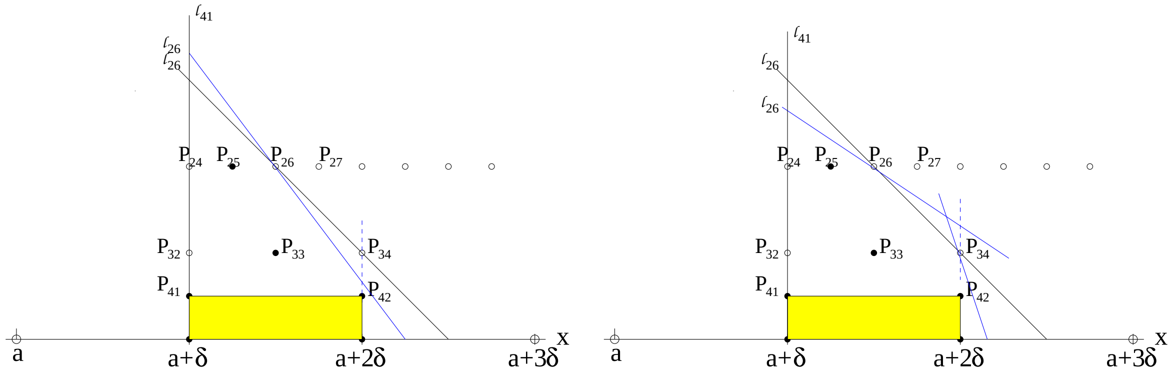


Figure 29: Subcase of Case 2-2, where $P_{25} \in \Omega'$. The slope of l_{41} is considered. It is clear that the larger the slope is, the larger the possible Ω' is.

If $P_{26} \in \Omega'$ (Figure 31), then P_{25} and P_{27} are not inside Ω'. Recall P_{41} is on the boundary. Three lines are critical, l_{25}, l_{27}, and l_{41}. Set Ω' is on the right of l_{25} and l_{27} and on the left of l_{41}. Clearly from the figures, in order to enclose a larger possible Ω', l_{27} should pass

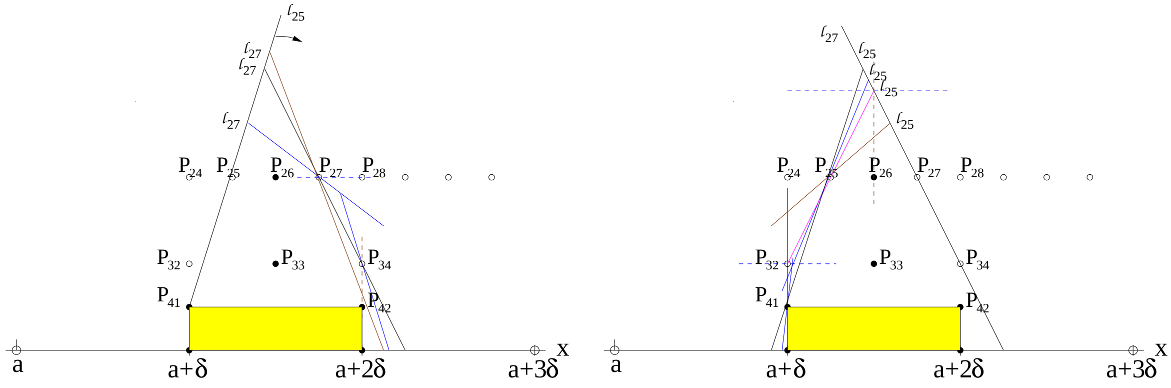


(a) P_{34} cannot be above l_{26} (b) P_{34} cannot be below l_{26}

Figure 30: Subcase of Case 2-2, where $P_{25} \in \Omega'$. The slope of l_{26} is considered. Comparing with the case where P_{34} is on l_{26}, (a) demonstrates that the maximal possible Ω' is smaller when P_{34} is above l_{26}; (b) demonstrates that the maximal possible Ω' is smaller when P_{34} is below l_{26}.

through P_{34} (Figure 31 (a)), and l_{25} and l_{27} are the same line (Figure 31 (b)). Therefore, the maximal possible Ω' is surrounded by l_{27} that passes throught P_{34}, line l_{41} that passes through P_{25}, and the x-axis. The relationship between the MER and the convex set Ω' is

$$\frac{|MER|}{|\Omega'|} \geq \frac{4}{17} > \frac{2}{9}.$$



(a) slope of l_{27} (b) slope of l_{25}

Figure 31: Subscase of case 2-2, where $P_{26} \in \Omega'$. Ω' is surrounded by l_{25}, l_{27} and l_{41}. (a) demonstrates that P_{34} cannot be below or above l_{27} in order to enclose a larger feasible Ω'. (b) demonstrates that P_{41} cannot be below or above l_{25} in order to have a larger feasible Ω'.

If $P_{27} \in \Omega'$ (Figure 32), Ω' are surround by lines l_{26}, l_{34} and l_{41}. Obviously, l_{34} is vertical to the x-axis. From the figure, when the slope of l_{41} is larger, larger Ω' could be enclosed.

The slope of l_{26} can be found not changing the maximal possible Ω'. Hence,

$$\frac{|MER|}{|\Omega'|_{\max}} = \frac{12}{49} > \frac{2}{9}.$$

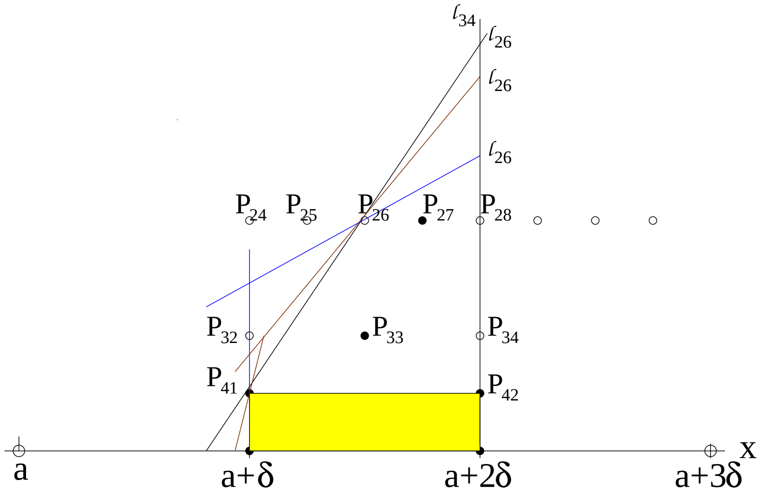


Figure 32: Subcase of Case 2-2, where $P_{27} \in \Omega'$. The maximal Ω' is limited by vertical l_{41}, vertical l_{34}, the x-axis, and l_{26}. Slope of l_{26} won't change the area of the maximal Ω' as long as it passes through P_{41}, or l_{41} is vertical.

If none of $P_{2j} \in \Omega'$, we can get more detailed subcases according to the status of P_{32}, P_{33} and P_{34}. In each case, however, it is easy to verify the area of the MER is at least $\frac{1}{4}$ of the area of Ω'. This number is greater than $\frac{2}{9}$. We leave this for the readers.

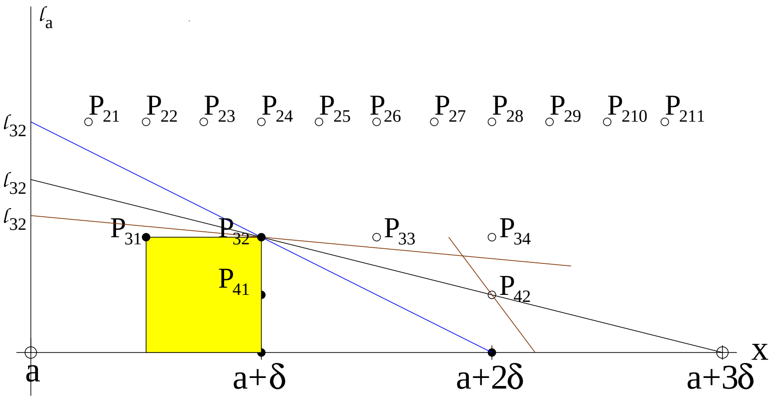
Hence, case C2-2 is proved.

3.3.2.6 Case C2-3

In case C2-3, the MER has vertices P_{31} and P_{32} and support $(a + 0.5\delta, a + \delta)$. Therefore, either P_{31} or P_{32} is on the boundary of Ω'.

If the point on the boundary is P_{32}, as shown in Figure 33, none of $P_{2x}, x \geq 1$ in Ω' since $(a+2\delta, 0)$ is inside Ω'. Hence, the limiting boundary for the maximal Ω' is: the vertical line passing $(a, 0)$ (l_a), l_{32} that passes P_{42}, and the x-axis. Please refer to Figure 33 for more details. Obviously, we have

$$\frac{|MER|}{|\Omega'|} \geq \frac{2}{9}.$$



l_a
l_{32}
l_{32}
l_{32}
P_{21} P_{22} P_{23} P_{24} P_{25} P_{26} P_{27} P_{28} P_{29} P_{210} P_{211}
P_{31}
P_{32}
P_{33}
P_{34}
P_{41}
P_{42}
x
a
a+δ
a+2δ
a+3δ

Figure 33: Subcase of Case C2-3, where P_{32} is on the boundary of Ω'. None of P_{2j} is inside of Ω'. Ω' is surrounded by the vertical line l_a, the x-axis, and l_{32} that also passes P_{42}.

On the other hand, if P_{31} is on the boundary of Ω' (Figure 34), notice $P_{33} \notin \Omega'$ and $(a + 2\delta, 0) \in \Omega'$. So, Ω' is between l_{31}, l_{33} and the x-axis. We can easily check that when the largest possible Ω' is enclosed, l_{33} passes through $(a + 2\delta, 0)$ since $P_{42} \notin \Omega'$, and l_{31} is vertical. Hence, in this case,

$$\frac{|MER|}{|\Omega'|} \geq \frac{2}{9}.$$

Therefore, case C2-3 is checked.

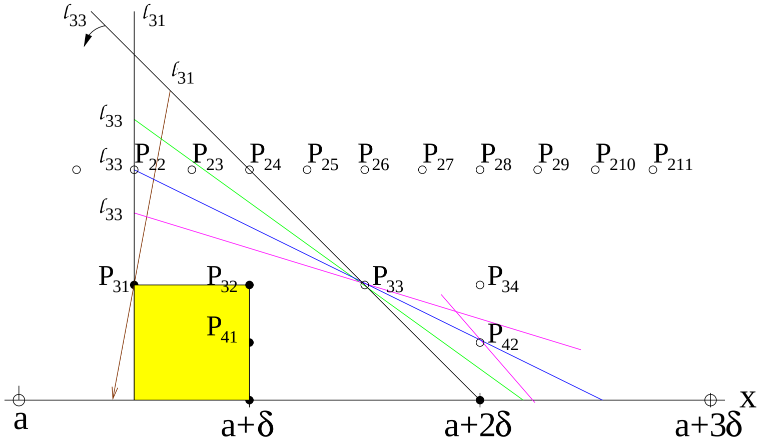


Figure 34: Subcase of Case C2-3, where P_{31} is on the boundary of Ω'. Region Ω' is between l_{31} and l_{33}. Note firstly, the larger the slope of l_{31}, the larger the possible Ω'. Secondly, $(a + 2\delta, 0)$ is below l_{33}. However, the larger the distance between them, the smaller the possible Ω'.

3.3.2.7 Case C2-4

For C2-4, either P_{32} or P_{33} is on the boundary of Ω'. The case while P_{33} is on the boundary is much easier than the other. We first consider the easier case. When P_{33} is on the boundary (Figure 35), $P_{31} \notin \Omega'$. Hence, Ω' is between l_{31} and l_{33}. For line l_{31}, obviously, the larger the slope of l_{31}, the larger the possible Ω'. Similar with the last subcase of C2-3, for l_{33}, , when l_{33} passes $(a + 2\delta, 0)$, the enclosed Ω' has larger area. Hence, we have

$$\frac{|MER|}{|\Omega'|} \geq \frac{2}{9}.$$

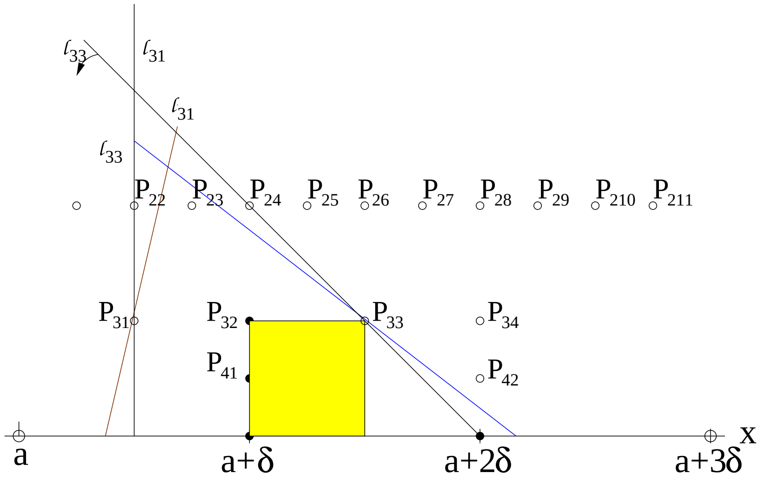


Figure 35: Subcase of Case C2-4, where P_{33} is on the boundary of Ω'. Region Ω' is between l_{31} and l_{33}. Note firstly, the larger the slope of l_{31}, the larger the possible Ω' is. Secondly, $(a + 2\delta, 0)$ is below l_{33}. However, the larger the distance between them, the smaller the possible Ω'.

Meanwhile, if P_{32} is on the boundary of Ω', we can find an l_{32} such that Ω' is on the right side of it. Furthermore, among P_{24} up to P_{27}, at most one of them will be in the Ω'. So we will have several subcases with respect to the status of each P_{2j}, $4 \leq j \leq 7$.

If $P_{24} \in \Omega'$ (Figure 36), then $P_{25} \notin \Omega'$. Furthermore, we have $P_{42} \notin \Omega'$. Hence, l_{32}, l_{25}, and l_{42} are crucial. Best states of these lines are: l_{32} is vertical to the x-axis, l_{25} is parallel to the x-axis, and l_{42} is vertical to the x-axis. Please refer to the figure. So,

$$\frac{|MER|}{|\Omega'|_{max}} > \frac{1}{4} > \frac{2}{9}.$$

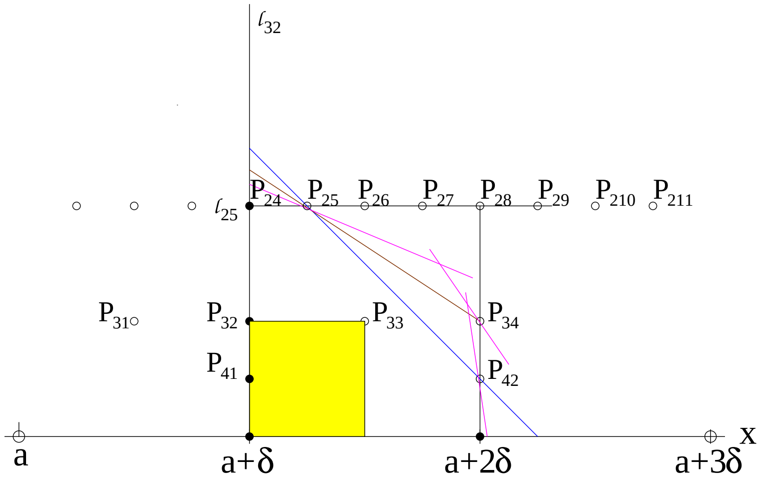


Figure 36: Subcase of Case C2-4, where P_{32} is on the boundary and $P_{24} \in \Omega'$. Ω' is bounded by l_{32}, l_{25}, l_{42}, and the x-axis. Line l_{32} is vertical because both P_{24} and P_{32} are on the line. l_{25} being zero and l_{42} being vertical will enclose larger Ω' that is applicable.

If $P_{25} \in \Omega'$ (Figure 37), Ω' is bounded by l_{32}, l_{26} and the x-axis. The best status is: l_{32} is vertical to the x-axis and l_{26} passes P_{42}, referring to the figure. Hence, we have

$$\frac{|MER|}{|\Omega'|} \geq \frac{12}{49} > \frac{2}{9}.$$

If $P_{26} \in \Omega'$ (Figure 38), Ω' is bounded by l_{32} and l_{27}. The best case, which includes a maximal possible Ω', is: l_{32} passing P_{25} and l_{27} passing P_{42}. Details can be found in the figure. Hence, we have

$$\frac{|MER|}{|\Omega'|} \geq \frac{15}{64} > \frac{2}{9}.$$

If $P_{27} \in \Omega'$ (Figure 39), Ω' is bounded by l_{32} and l_{28}. The best case is: l_{32} passing P_{26} and l_{28} vertical. Details can be found in the figure. Again, we have

$$\frac{|MER|}{|\Omega'|} \geq \frac{2}{9}.$$

If none of $P_{2x} \in \Omega', x = 4, 5, 6, 7$ (Figure 40), Ω' could be surrounded by l_{32}, l_{24}, and l_{28}. One of the best status is: l_{32} is vertical to the x-axis, l_{25} is parallel to the x-axis, and l_{42} is

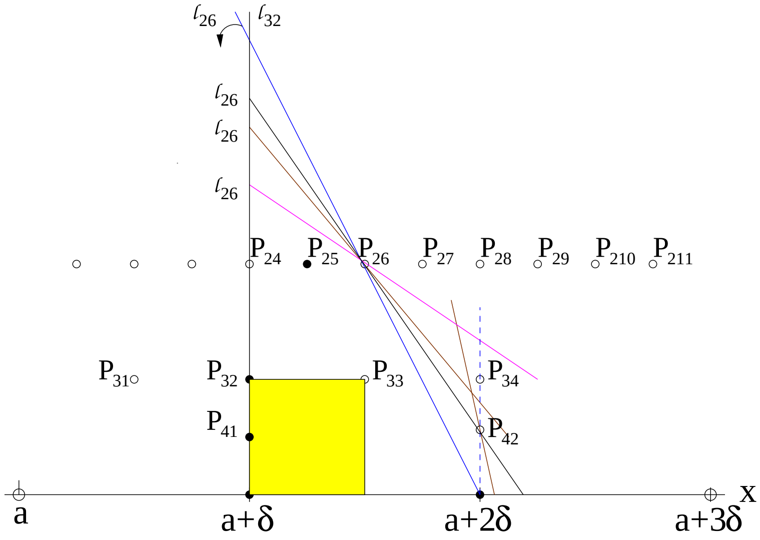
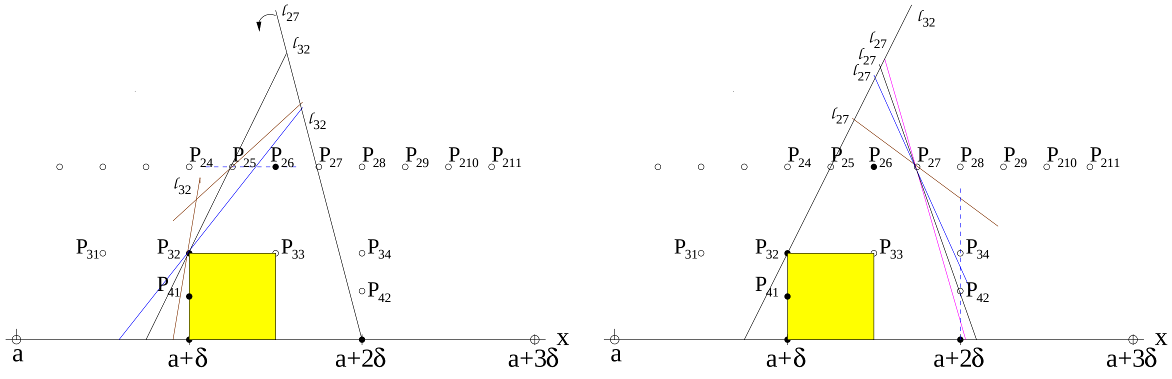


Figure 37: Subcase of Case C2-4, where P_{32} is on the boundary of Ω' and $P_{25} \in \Omega'$. Ω' is bounded by l_{32}, l_{26}, and the x-axis. For l_{32}, the larger the slope is, the larger the possible enclosed Ω' is. For l_{26}, point $(a + 2\delta, 0)$ is below l_{26}. The larger the distance between them is, the larger the possible Ω' is. Meanwhile, l_{26} is below P_{42} since P_{42} not in Ω'. Hence, the best l_{26} passes P_{42}.



(a) slope of l_{32} (b)

slope of l_{27}

Figure 38: Subcase of Case C2-2, where P_{32} is on the boundary and $P_{26} \in \Omega'$. Ω' is bounded by l_{32} and l_{27}. (a) indicates that in order to include a larger possible Ω', l_{32} should pass P_{25}, or more correctly, slightly below P_{25}; (b) indicates that l_{27} should pass P_{42} to include larger area.

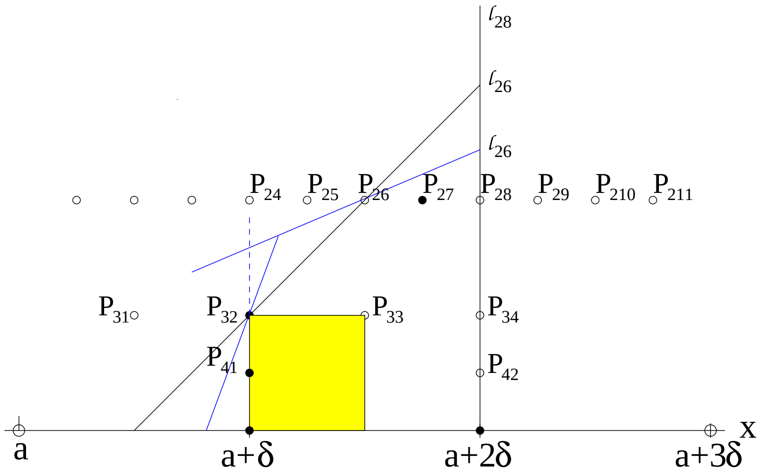


Figure 39: Subcase of Case 2-4, where P_{32} is on the boundary and $P_{27} \in \Omega'$. Region Ω' is bounded by l_{32} and l_{28}. For l_{32}, when P_{26} is on it, the embedded area is larger than the area when P_{26} is not on the line. Line l_{28} should be vertical because P_{34} and P_{42} are not in Ω'.

vertical to the x-axis. Hence,

$$\frac{|MER|}{|\Omega'|} > \frac{1}{4} > \frac{2}{9}.$$

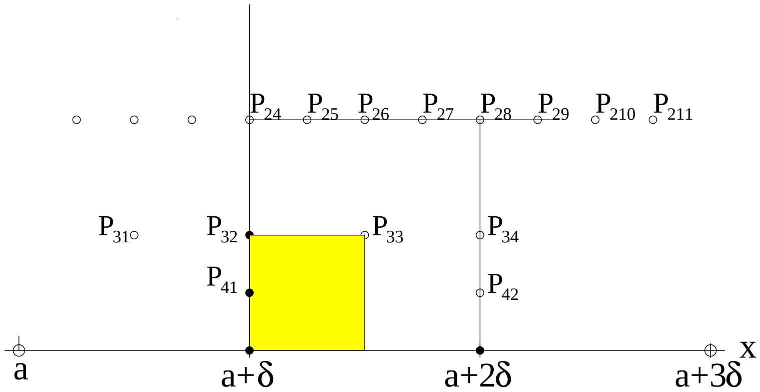
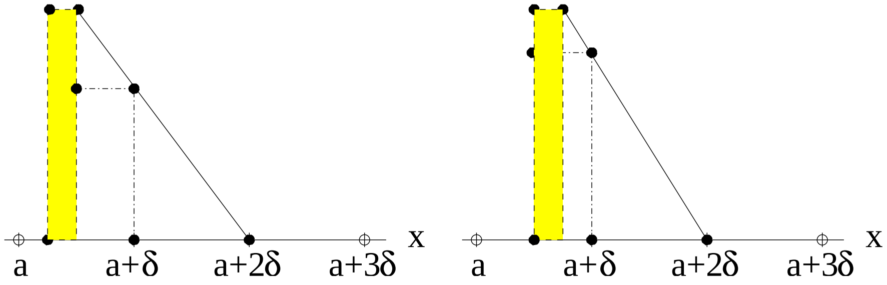


Figure 40: Subcase of Case C2-4, where P_{32} is on the boundary of Ω' and none of $P_{2j} \in \Omega'$. Ω' is surrounded by l_{32}, l_{24}, and l_{28}. Line l_{28} is vertical because P_{34} is outside Ω' and $(a + 2\delta, 0)$ is inside. l_{24} is horizontal because l_{28} is vertical and more area are supposed to be enclosed. Hence, given l_{28} and l_{24}, the slope of l_{32} can be any positive value as long as $(a, 0)$ is above it.

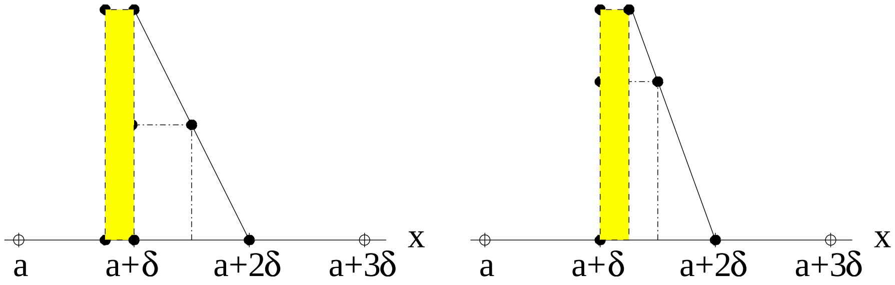
3.3.2.8 Cases C2-5, C2-6, C2-7, C2-8, and C2-9

Cases C2-5 to C2-8 are similar to cases C1-3 and C1-4 (Figure 41 and Figure 42), where another embedded dyadic rectangle with larger area or with the same area but longer support can be found. Hence, these cases are either impossible or covered by other cases.



(a) C2-5 (cannot happen) (b) C2-6 (cannot happen)

Figure 41: Case C2-5 & Case C2-6. Two impossible cases because a larger embedded dyadic rectangle can be found with support $(a + 0.5\delta, a + \delta)$.



(a) C2-7 (covered by another case) (b) C2-8 (cannot happen)

Figure 42: Case C2-7 & Case C2-8. (a) indicates that case C2-7 is covered by another case since comparing with the shaded part, a lower embedded rectangle with the same area but smaller height can be found; (b) indicates that case C2-8 is impossible because a larger dyadic rectangle can be found.

3.3.2.9 Case C2-9

This case is almost the same as case C1-5. For point P_{25} and point P_{26}, at least one of them should be on the boundary of Ω'. We first assume that point P_{26} is on the boundary (Figure 43 and Figure 44). Hence, there exists a line l_{26} such that Ω' is on the left side of this line. Furthermore, we have $P_{32} \notin \Omega'$, so there is a line l_{32} such that Ω' is on the right side of it. The best choice for such l_{32} and l_{26} is that l_{32} is vertical (Figure 43) and l_{26} passing P_{34}. Therefore, we have

$$\frac{|MER|}{|\Omega'|} \geq \frac{2}{9}.$$

On the other hand, when P_{25} is on the boundary of Ω' (Figure 45 and Figure 46). Lines

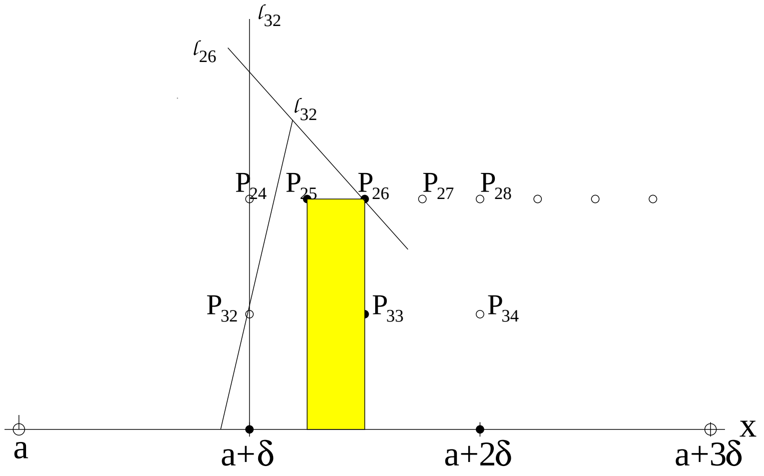
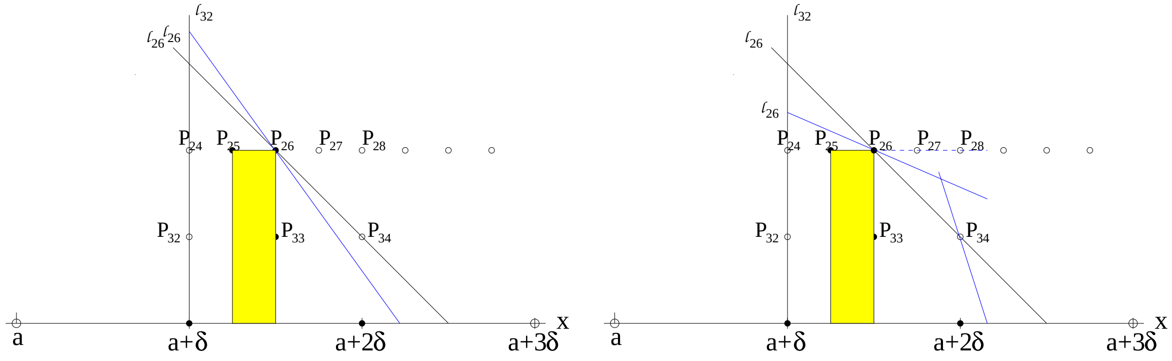


Figure 43: Subcase of Case C2-9, where P_{26} is on the boundary of Ω'. Region Ω' is between l_{32} and l_{26}. The slope of l_{32} is considered here. The line should be vertical such that the enclosed area is larger.



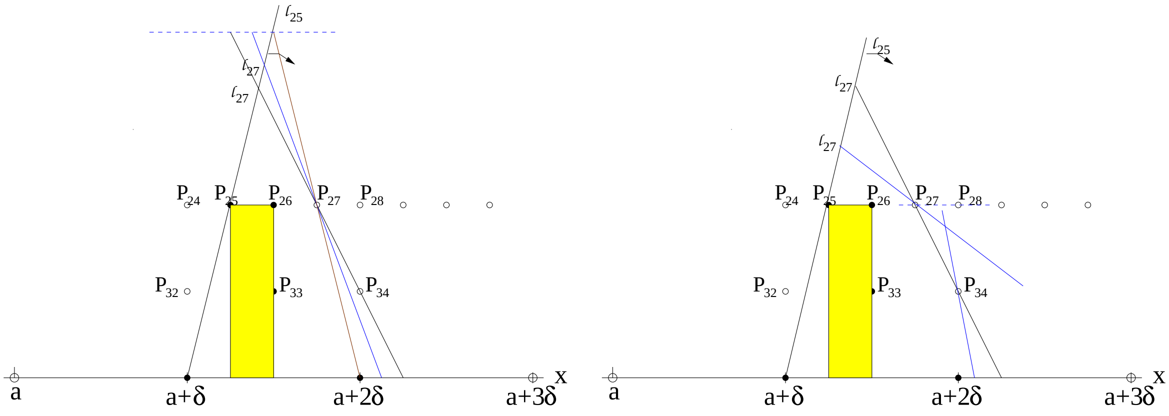
(a) P_{34} cannot be above l_{26}

(b) P_{34} cannot be below l_{26}

Figure 44: Subcase of Case 2-9, where P_{26} is on the boundary of Ω'. Region Ω' is between l_{32} and l_{26}. The slope of l_{26} is considered here. In order to include more area, (a) indicates that P_{34} cannot be above l_{26} and (b) indicates that P_{34} cannot be below l_{26}.

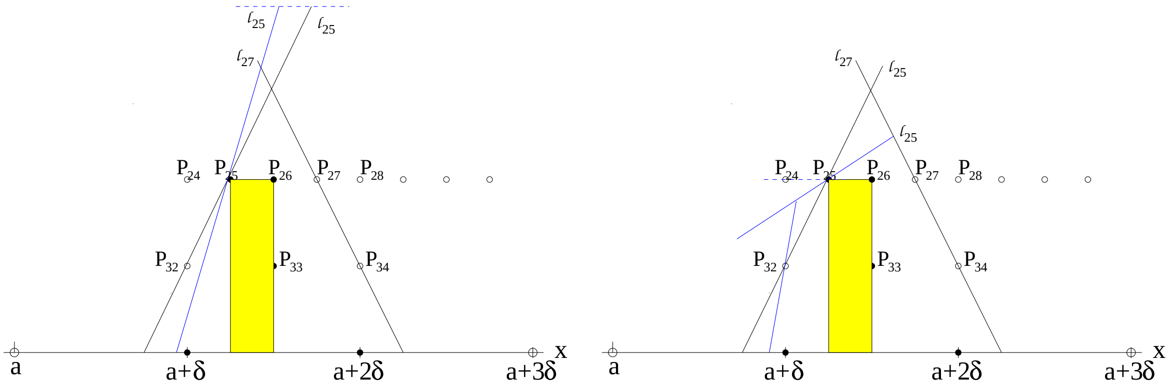
l_{27} and l_{25} are crucial, because Ω' is between them. From those two figures, the enclosed possible Ω' is maximized if l_{27} passes P_{34} (Figure 45) and l_{25} passes P_{32} (Figure 46). Hence, we have

$$\frac{|MER|}{|\Omega'|} \geq \frac{2}{9}.$$



(a) P_{34} cannot be above l_{27} (b) P_{34} cannot be below l_{27}

Figure 45: Subcase of Case 2-9, where P_{25} is on the boundary. Consider the slope of l_{27}, (a) indicates that the enclosed area is smaller if P_{34} is above l_{27}; (b) indicates that the enclosed area is smaller if P_{34} is below l_{27}.



(a) P_{32} cannot be above l_{25} (b) P_{32} cannot be below l_{25}

Figure 46: Subcase of Case 2-9, where P_{25} is on the boundary. We consider the slope of l_{25}. (a) indicates that the enclosed area is smaller if P_{32} is above l_{25}; (b) indicates that the enclosed area is smaller if P_{32} is below l_{25}.

Based on all the above, we have proved the Theorem 2.1.

3.4 Discussion and Conclusion

Similar works, regarding constants related to multiscale methods in other scenarios, can be found in [47, 48].

Note for Theorem 2.1, there may exist other partitions or other methods, which could give a simpler proof. But our intention is to show that such a non-zero constant, 2/9, exists. This constant tells us that as long as h(v)-parallelograms are detectable, convex sets are detectable. Hence we can guarantee the detectability of convex sets. Such a result can be used in the pre-screening of large volume of images. We have stated several potential applications in Introduction.

More discussions and potential applications will be provided in [74], which is a derivative of this part of the thesis. We will try to exam the results in real problems, such as the Cryo-EM images.

PART II

Regression Models

CHAPTER IV

INTRODUCTION

4.1 Linear Model

Linear regression is one of the most widely used statistical technique for investigating the relationship between variables. Applications of linear regression are numerous and occur in almost every field, including engineering, medical science, economics, psychology, management, and many more. It has been a mainstay of statistics for the past decades and remains one of the most important tools.

The linear regression model assumes that the relationship between the expected response (denoted by $E(Y|A)$) and the predictors (denoted by $A_j, j = 1, 2, ..., m$) is linear, or can be reasonably approximated by a linear model. Mathematically, the linear regression model is

$$Y = A\mathbf{x} + \varepsilon,$$

where the notations are explained as follows:

- $Y = [Y_1, Y_2, \cdots, Y_n]^T$ is the response vector, in which Y_i is the observed response in the ith trial;
- $A = [A_1, A_2, \cdots, A_m] \in \mathbf{R}^{n \times m}$ is called model matrix, in which $A_j = [a_{1j}, a_{2j}, \cdots, a_{nj}]^T$, $j = 1, 2, ..., m$, is the value of the jth predictor in all trials;
- $\mathbf{x} = [x_1, x_2, ..., x_m]^T$ are unknown parameters (or coefficients) that we want to estimate;
- $\varepsilon = [\varepsilon_1, \varepsilon_2, ..., \varepsilon_n]^T$ are the random errors, which are sampled from distribution F with mean zero. Traditionally, people assume that ε_i's are i.i.d. normal distributed.

Linear models were largely developed in the pre-computer age of statistics, because they are simple and easy to be interpreted. However, even in today's computer era there are still good reasons to study and use the linear models due to several reasons. First of all, although complicated non-linear models are available right now, for predictions purposes, linear models can sometimes outperform fancier models, especially in situations with small numbers of training cases or sparse data. Secondly, notice that the linear model is linear in the parameters, not the variates. The variates A_j can be quantitative inputs, transformations of quantitative inputs, qualitative values, or interaction between variables. This expansion considerably extends the scope of the linear regression. Thirdly, real world is much more complicated than the theoretical assumption. In many applications, errors are not normal distributed, inputs can be correlated, or data could be misreported. Developing robust estimators that can survive the distortion is an interesting problem. Finally, because of the rapid development of the computer resources, the size of the data for analysis becomes much larger. Some of the data contain thousands of variates. It's impossible and unpractical to interpret the model with huge number of predictors. How to eliminate the low-effect variables and contains the most related ones is another interesting problem.

Following the above concerns, two aspects have interested us.

- One is the robustness of the model. The word "robust" in this chapter means the insensitivity against the error distributions that belong to a family, in which the probability of large errors is small however present. How can we develop a meaningful estimator that can remove or reduce the effect of large errors? How is it compared with the traditional robust estimators such as M-, L-, and R- estimators?

- The other is the sparsity. The word "sparsity" means that the number of predictors useful for the prediction or explanation are significantly less than the total number of predictors. Fast algorithms are introduced for estimation and variable selections. Statistical criteria are developed to guide the model selection through many considerations. Some interesting questions are: how well a greedy algorithm is used in subset selection? How can greedy approaches be connected with global statistical criterion of optimality?

We will try to analyze and answer these questions in the second half of this dissertation.

4.2 Contributions

This part of this dissertation is to give new results with respect to the linear model. As mentioned in the previous section, two aspects of the linear models are studied: regression with non-Gaussion noise and variable selection through stepwise and/or all-subset selection algorithms. Specifically, we have developed new regression mechanism for noise with outliers, analyzed the performance of certain stepwise algorithms in subset selection, and proposed a new all-subset selection algorithms. The main contribution of this part can be summarized as follows:

- We have derived a new robust estimator appropriate for the linear regression model with stochastically bounded noise. Given this type of noise and some necessary regularity conditions, we show that this robust estimator is a locally asymptotical minimax estimator. Simulations on the real as well as artificial data demonstrate the advantages of this new estimator over the Least Square Estimator and the Huber's M-estimator. An easy-to-implement algorithm is obtained based on the proximal point method. We also present an alternative approach, using a state-of-the-art optimization software, to solve the derived estimation problem.

- We analyze the effectiveness of least angle regression in correctly retrieving the original variables that produce the signal. We revisit the connection between least angle regression and Lasso by showing that least angle regressions give the same solution as Lasso. We also provide a counterexample in which least angle regressions cannot get the correct subset. This counterexample stirs the interest in finding the condition for accurate model selection. We prove that many existing criteria in subset selection means to solve an NP-hard problem. But its solution, under some conditions, is the same with the solution that certain stepwise algorithms provide. We study the these conditions that leads to common solutions. Several conditions are derived, from different aspects.

- We also study the all subset searching algorithms for linear model. The leaps-and-bounds algorithm is currently the state-of-the-art. We review the algorithm with our new data structure, which is easier to be understood than the original description. Based on the same framework, we introduce an enhanced algorithm by including newly designed optimality tests in each iteration in order to exclude (i.e., leap) more non-optimal subsets. Simulations validate the improvements.

4.3 Organization of part II

The rest of part II is organized as follows.

- Chapter 5 derives RSBN (regression with stochastically bounded noises) as a new robust estimator. It is proven that RSBN estimator is a locally asymptotic minimax estimator. The derived estimator is compared with the least square estimator, which is a mainstay of statistics, and the Huber's estimator, which heavily influenced the development of robust estimators.

- Chapter 6 focuses on the concurrence of two optimization solutions. One is the solutions of certain stepwise algorithms, such as LARS for Lasso. The other is the existing criteria in subset selection. It is shown that in some cases, these two problems can have concurrent solutions. We derive several conditions for the exact recovery for either problem, and for both of them. An extreme example with respect to the least angle regressions is constructed, which by itself is interesting.

- Chapter 7 develops an advanced algorithm for all subset selection, based on the leaps-and-bounds algorithm by Furnival and Wilson (1974). New optimality tests are added into the original simple tests that based only on the the residual sums of squares. The new method brings more information under consideration, so that it can exclude more non-optimal subsets.

- Appendix contains the details of some proofs. Appendix A gives the proofs associated with the development of RSBN. Appendix B gives the proofs regarding the results of LARS.

CHAPTER V

RSBN: REGRESSION WITH STOCHASTICALLY BOUNDED NOISES

In this chapter, we consider M-estimates in a regression model where the noises are of unknown but stochastically bounded distribution. An asymptotic minimax M-estimate is derived. Simulations demonstrate the robustness of this approach, as well as advantages over commonly used estimates such as the ordinary least square estimate and the Huber's estimate. The new method is named *regression with stochastically bounded noises* (RSBN). We provide an iterative numerical solution, which is derived from the proximal point method. The iterative method is elegant, however may not have fast rate of convergence. RSBN can also be solved by applying existing state-of-the-art nonlinear optimization software. We present SNOPT as one example. Insights from RSBN are discussed.

This chapter is organized as follows. Section 5.1 summarizes the contributions of this chapter. Section 5.2 presents the formulation and the main theoretical result. Section 5.3 establishes the asymptotic minimaxity of the proposed estimate. Section 5.4 describes the numerical algorithm that is derived from the proximal point method. Related analysis on the convergence of this algorithm is presented. Section 5.5 presents an alternative numerical approach, which utilizes a state-of-the-art but commercialized optimization software. Section 5.6 conducts simulations that consolidate our findings. Section 5.7 and Section 5.8 present the discussions and the conclusions, respectively.

5.1 Introduction

We consider a regression problem in which the noise distribution is unknown, but some probabilistic information is available. More specifically, we consider the cases when the noise is stochastically bounded: there exist constants $\delta > 0$ and $0 < \alpha < 1$, such that Prob.$\{|\text{noise}| > \delta\} < \alpha$. In a regression framework, we derive the asymptotic minimax estimate of the coefficients for all noise distributions satisfying the above condition.

Interesting similarity between the derived minimax estimate and some recently emerged criterion functions in model selection is inspiring. Specifically, the fact that the objective function become linear outside a neighborhood of the origin coincides with the ℓ_1-norm principle that has recently gained popularity via methods such as Lasso [90] and Basis Pursuit [12].

RSBN can be viewed as an extension of the well-developed Huber M-estimate. Hence it is a development in the line of robust statistics. We found that by deriving the exact form of the asymptotic minimax estimate of the coefficients, we can achieve slightly better numerical performance. Simulations on synthetic data are reported to demonstrate our findings.

Using the proximal point method in optimization, we develop an iterative approach that is extremely simple to implement — it takes a few lines in MATLAB. However, its numerical performance is not satisfactory: it can converge extremely fast in some situations, and extremely slow in some pathological cases. We give our analysis on the speed of convergence in some simplified situations. We also present an alternative: using existing state-of-the-art optimization software packages, e.g., SNOPT.

5.2 Formulation and Main Theoretical Result

Recall that a regression model is

$$\mathbf{y} = A\mathbf{x} + \varepsilon,$$

where $\mathbf{y} = (y_1, y_2, \cdots, y_n)^T \in \mathbf{R}^n$ is the response vector, $\mathbf{x} \in \mathbf{R}^m$ is a vector of coefficients, model matrix is $A = [a_1, a_2, \cdots, a_n]^T \in \mathbf{R}^{n\times m}$, and a random error vector is $\varepsilon = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)^T$. Without loss of generality, for the rest of the chapter, we assume that vectors a_i's are standardized (i.e., $\|a_i\|_2 = 1,$ for $i = 1, 2, \ldots, n$) and the model matrix A is of full column rank (equivalently, matrix inverse $(A^TA)^{-1}$ exists). Furthermore, we assume that the random errors $\varepsilon_i, i = 1, 2, \ldots, n$, are i.i.d. with a common density function f.

Given a set of coefficients $\mathbf{x}$, the residual associated with the ith response is $r_i = y_i - a_i^T\mathbf{x}$.

One can estimate the set of coefficients by solving the following optimization problem:

$$\begin{aligned} \text{minimize} \quad & \textstyle\sum_{i=1}^{n} \rho(r_i), \\ \text{subject to} \quad & r_i = y_i - a_i^T \mathbf{x}, \quad i = 1, 2, \cdots, n. \end{aligned} \tag{7}$$

Here, we normally require function ρ to be convex; because convex optimization problem in principle is much more amenable than other optimization problems (e.g., combinatoric optimization problems). If we define a residual vector $\mathbf{r} = (r_1, r_2, \cdots, r_n)^T \in \mathbf{R}^n$, the restriction of the above optimization problem can be rewritten as $\mathbf{r} = \mathbf{y} - A\mathbf{x}$. Another way to express the optimization problem in (7) is:

$$\begin{aligned} \text{minimize} \quad & \rho(\mathbf{r}) = \textstyle\sum_{i=1}^{n} \rho(r_i), \\ \text{subject to} \quad & \mathbf{r} = \mathbf{y} - A\mathbf{x}. \end{aligned}$$

A key feature of the above formulation is that the criterion function (which is also the objective) is an additive function with respect to the residuals r_i. The criterion depicted in (7) covers many known approaches. For example, when $\rho(x) = x^2$, we have the ordinary least square estimate.

We consider the situation when the random errors ε_i satisfy the following condition.

Condition 2.1 (stochastically bounded noises) *In a regression model, if for i.i.d. random errors $\varepsilon_i, i = 1, 2, \ldots, n$, we have*

$$\textit{Prob.}(|\varepsilon_i| > \delta) \leq \alpha, \quad \forall 1 \leq i \leq n,$$

where $\delta > 0$ and $0 < \alpha < 1$ are predetermined, then we have stochastically bounded noises.

In this chapter, we propose the following function for $\rho(x)$:

$$\rho(x) = \begin{cases} -\log\cos\lambda_1(x/\delta), & \text{if } |x/\delta| < 1; \\ \lambda_1 \tan\lambda_1 \cdot |x/\delta| - \lambda_1 \tan\lambda_1 - \log\cos\lambda_1, & \text{if } |x/\delta| \geq 1. \end{cases} \tag{8}$$

where $0 < \lambda_1 < \pi/2$ is a function of α. The analytic relation between λ_1 and α will be established when we derive the asymptotic minimaxity of the above estimate. Figure 47 gives a graphical comparison between the above ρ and the objective functions that are used in the least square estimate and the Huber's M-estimate.

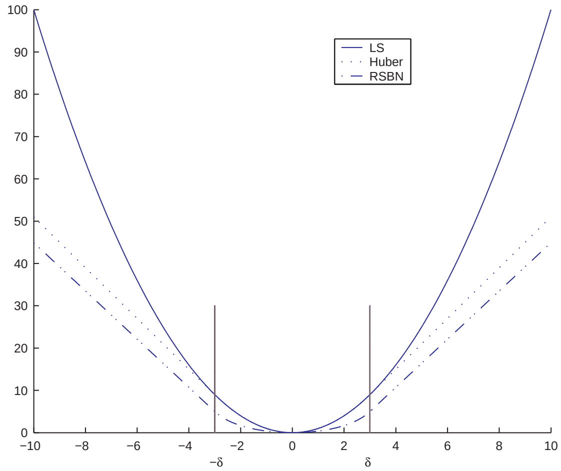


Figure 47: Objective function $\rho(x)$ in ordinary least square, Huber's M-estimate, and RSBN.

When the function $\rho(x)$ has the form in (8), the obtained estimate is called a *regression with stochastically bounded noise* (RSBN) estimate. With our choice of ρ, problem (7) turns into a nonlinear optimization problem. The main reason to choose the function ρ in (8) is the following theorem.

Theorem 2.2 *Under the 'stochastically bounded noises' condition, the estimate from (7) with the function ρ specified in (8) is the asymptotic local minimax estimate of the coefficient vector* $\mathbf{x}$*.*

The above theorem will be established in the next section. Note we proved local minimaxity, instead of global minimaxity. Distinction between the two will be discussed in Section 5.7.5.

5.3 Regression Achieving Asymptotic Minimaxity

Theoretical foundation of RSBN will be presented in the following subsections:

- Asymptotic normality (Section 5.3.1): we establish that the solution to (7) is asymptotically normal.
- Minimum asymptotic variance estimation (Section5.3.2): we derive the estimate that achieves the minimum asymptotic variance.
- Least informative distribution (Section 5.3.3): we study the worst case in estimation,

which is equivalent to finding the least informative distribution. By doing so, we get a locally asymptotic minimax estimate.

- Regression with stochastically bounded noises (RSBN) (Section 5.3.4 and 5.3.5): we present our regression method, by specifying the function $\rho(\cdot)$ in (7).
- Fisher information (Section 5.3.6) and asymptotic variance (Section 5.3.7): we derive the Fisher information for the least informative distribution and the asymptotic variance for the RSBN estimate.
- Robustness (Section 5.3.8): we consider the robustness of the estimate by specifying its breakdown point.

5.3.1 Asymptotic Normality

The solution to (7) is an M-estimate. In this section, we derive the asymptotic normality of an M-estimate.

We start with assumptions and notations. First, we consider location estimation. In (7), we temporarily assume that $m = 1$ and $a_i = 1, i = 1, 2, \cdots, n$. Suppose ρ has the second derivative. Let $\psi = \rho'$ be the first derivative of ρ. Define a function $\lambda(t, F) = \int \psi(\xi - t) dF(\xi)$, for $t \in \mathbf{R}$, where F is a cumulative distribution function (c.d.f.) of a random variable ξ. Define a functional $\mathbf{T}$ from distribution space to $\mathbf{R}$, such that $\lambda(\mathbf{T}(F), F) = 0$. Value $\mathbf{T}(F)$ is defined as the true location parameter. Let F_n be the empirical c.d.f. Note that $\lambda(x, F_n) = 0$ is the first order necessary condition (FOC) for a minimizer of (7). It is easy to see that $\mathbf{T}(F_n)$, which satisfies $\lambda(\mathbf{T}(F_n), F_n) = 0$, is an M-estimate for n samples.

The asymptotic normality theorem is typically derived in the following three steps:

1. Firstly, we have

$$\begin{aligned} 0 &= \lambda(\mathbf{T}(F), F) - \lambda(\mathbf{T}(F_n), F) + \lambda(\mathbf{T}(F_n), F) - \lambda(\mathbf{T}(F_n), F_n) \\ &= [\mathbf{T}(F) - \mathbf{T}(F_n)] \frac{\lambda(\mathbf{T}(F), F) - \lambda(\mathbf{T}(F_n), F)}{\mathbf{T}(F) - \mathbf{T}(F_n)} \\ &\quad - \frac{1}{n} \sum_{i=1}^{n} \left[\psi(y_i - \mathbf{T}(F_n)) - \lambda(\mathbf{T}(F_n), F) \right]. \end{aligned} \tag{9}$$

2. We assume some regularity conditions are satisfied, and $\mathbf{T}(F_n) \to \mathbf{T}(F)$. (As long as ψ is monotone and $F_n \Rightarrow F$, which are generally satisfied conditions, $\mathbf{T}(F_n) \to \mathbf{T}(F)$ is true.) We have

$$\begin{aligned} \frac{\lambda(\mathbf{T}(F), F) - \lambda(\mathbf{T}(F_n), F)}{\mathbf{T}(F) - \mathbf{T}(F_n)} &\Rightarrow \frac{\partial}{\partial t}\lambda(t, F)|_{t=\mathbf{T}(F)} \\ &= \int \psi'(x - \mathbf{T}(F))dF(x). \qquad (10) \end{aligned}$$

Since ρ has the second derivative, the derivative ψ' exists. The above also implies that the right hand side of (10) is integrable.

3. Observe

$$\begin{aligned} &\frac{1}{n}\sum_{i=1}^{n}\left[\psi(y_i - \mathbf{T}(F_n)) - \lambda(\mathbf{T}(F_n), F)\right] \\ \Rightarrow &\frac{1}{\sqrt{n}}\text{Normal}\left(0, \int \psi^2(x - \mathbf{T}(F))dF(x)\right). \qquad (11) \end{aligned}$$

This is a direct result from central limit theorem (CLT) because the left hand side is a sum of i.i.d. random variables. We suppose to check the Lindeberg condition. In this chapter, we assume the condition is satisfied. For more details, see [38].

Combining (9), (10) and (11), we have

$$\mathbf{T}(F_n) - \mathbf{T}(F) \sim \frac{1}{\sqrt{n}}\text{Normal}\left(0, \frac{\int \psi^2 dF}{(\int \psi' dF)^2}\right).$$

The asymptotic variance is equal to $\frac{\int \psi^2 dF}{(\int \psi' dF)^2}$.

The above result can be generalized to a multivariate parameter case. When $m > 1$ and matrix A is of full column rank, the asymptotic variance/covariance matrix of the M-estimate will be $\frac{\int \psi^2 dF}{(\int \psi' dF)^2}(A^T A)^{-1}$. For reference, please see Chapter 7.6 in [44].

Lemma 3.1 *Given function $\rho(\cdot)$ that has a monotone increasing first derivative $\psi = \rho'$ and its second derivative is integrable in (10), the estimate given by (7) has the asymptotic distribution*

$$\textit{Normal}\left(\mathbf{x}_0, \frac{1}{n}\frac{\int \psi^2 dF}{(\int \psi' dF)^2}(A^T A)^{-1}\right),$$

where the vector $\mathbf{x}_0$ is made of the true values of the coefficients.

We take the quantity $\frac{\int \psi^2 dF}{(\int \psi' dF)^2}$ as a natural measure of performance for an M-estimate. The smaller this quantity, the closer the M-estimate is to the true value of the parameter.

5.3.2 Minimum Asymptotic Variance Estimation

We call the quantity $\frac{\int \psi^2 dF}{(\int \psi' dF)^2}$ the *asymptotic variance*. It is known that the asymptotic variance is lower bounded by the inverse of Fisher information. The following analysis is well-adopted in mathematical statistics.

Let $f_\theta = f(x-\theta)$ be the p.d.f. associated with c.d.f. F_θ and location parameter θ. $I(f)$ is the Fisher information with respect to θ. We have

$$\lambda(\theta, F_\theta) = \int \psi(x-\theta) f(x-\theta) dx = \text{constant}.$$

Taking the operator $\frac{\partial}{\partial\theta}$ on both sides, we get

$$0 = -\int \psi'(x-\theta) f(x-\theta) dx - \int \psi(x-\theta) f'(x-\theta) dx. \tag{12}$$

Here we assume both ψ and f are absolutely continuous and have first derivatives. From (12),

$$\begin{aligned} 1 &= \left[\int \left(\frac{\psi}{\int \psi' f}\right) \cdot \left(-\frac{f'}{f}\right) f\right]^2 \\ &\overset{\text{Cauchy}}{\leq} \int \left(\frac{\psi}{\int \psi' f}\right)^2 f \cdot \int \left(-\frac{f'}{f}\right)^2 f \\ &= \int \left(\frac{\psi}{\int \psi' f}\right)^2 f \cdot I(f), \end{aligned}$$

where $I(f)$ is the Fisher information of f. So asymptotic variance $\int \left(\frac{\psi}{\int \psi' f}\right)^2 f \geq \frac{1}{I(f)}$.

It achieves equality iff $\rho' = \psi \propto -\frac{f'}{f} = (-\log f)'$, in which case the M-estimate is also the maximum likelihood estimate (MLE). When $\rho = -\log f$, we call the solution to (7) the *minimum asymptotic variance estimate*. The result in this subsection is summarized as the following lemma.

Lemma 3.2 *The asymptotic variance of the estimate from (7) is lower bounded by* $1/I(f)$*. The lower bound is achieved when* $\rho \propto (-\log f)$*, i.e., when the estimate is the maximum likelihood estimate.*

5.3.3 Least Informative Distribution

The smaller the Fisher information $I(f)$ is, the larger is the lower bound of the asymptotic variance. We are interested in the least informative distribution, which is the solution to the following optimization problem: (note the variable is a function f)

$$\begin{aligned} \text{minimize} \quad & I(f), \\ \text{subject to} \quad & \textstyle\int v(x)f(x)dx \leq 0, \\ & \textstyle\int f(x)dx = 1. \end{aligned} \tag{13}$$

Note in our framework, function f is assumed to have second derivative. Otherwise, a piecewise constant function f may lead to $I(f) = 0$, which leads to infinite asymptotic variance. Such a case is excluded by demanding the existence of the second derivative.

The first constraint is a general form of many types of restrictions on the noise distribution. For example, if

$$v(x) = \begin{cases} -\alpha, & |x| < \delta, \\ 1 - \alpha, & |x| \geq \delta, \end{cases} \tag{14}$$

we have $\int_{-\delta}^{\delta} f \geq 1 - \alpha$. This implies stochastically bounded noises. This condition is meaningful when there are outliers. If $v(x) = x^2 - B$, we have $\int x^2 f(x)dx \leq B$, which is the second moment constraint. Similarly, we can have some other moments constraints. The second constraint in (13) is the constraint of a p.d.f.

To find the solution to (13), we consider the following function:

$$\mu(f) = I(f) + \beta_1[\int v(x)f(x)dx + \gamma^2] + \beta_2[\int f(x)dx - 1],$$

where β_1 and β_2 are the Lagrange multipliers, and $\gamma \in \mathbf{R}$ is a pseudo-variable: $\int v(x)f(x)dx + \gamma^2 = 0$. We consider a variational approach. Assume function f_0 is a minimizer in (13). For any other p.d.f. f_1, consider $f_t = (1-t)f_0 + tf_1$, $0 \leq t \leq 1$. Because f_0 is a minimizer, we must have $\frac{d}{dt}\mu(f_t)|_{t=0} \geq 0$ for any f_1, which is equivalent to

$$-4\int \frac{(\sqrt{f_0})''}{\sqrt{f_0}}(f_1 - f_0)dx + \beta_1 \int v \cdot (f_1 - f_0)dx + \beta_2 \int (f_1 - f_0)dx \geq 0.$$

The above holds if and only if

$$4\frac{(\sqrt{f_0})''}{\sqrt{f_0}} - \beta_1 \cdot v - \beta_2 = 0. \tag{15}$$

Note the above is a necessary condition for f_0 to be the solution to (13).

Lemma 3.3 *If a function f_0 has second derivative and achieves a local minimum in (13), then it satisfies the equation (15).*

In the next subsection, we construct a function f_0 that satisfies (15). This constructed function f_0 leads to the objective function that is used in our RSBN.

5.3.4 Regression with Stochastically Bounded Noises (RSBN)

Recall our objective is to find an appropriate function ρ in (7), so that the solution to (7) is both easy to compute and optimal within a family of distributions for random errors.

In our construction, the following conditions are satisfied.

- [Conditions for probability density function] Function f is a probability density function. Function f is from real numbers to nonnegative real numbers $f : \mathbf{R} \to \mathbf{R}^+$ ($f \geq 0$) and $\int f = 1$. In previous discussion, we implied that function f has finite Fisher information, $I(f) < \infty$. We also assume that the density function f is symmetric about 0.

- [Conditions for stochastically bounded noises] We have $\int_{-\delta}^{\delta} f \geq 1 - \alpha$. This means that the probability of noises having absolute values no larger than δ is at least $1 - \alpha$. Usually α is small. It is equivalent to say that no more than proportion α of residuals can have absolute values greater than δ. As mentioned earlier, an equivalent expression of this condition is $\int v(x)f(x)dx \leq 0$, where function v is defined in (14).

- [Conditions for convexity] The function $\rho(x) = -\log f(x)$ must be convex, otherwise we will not have a convex optimization problem. The first derivative of ρ, ρ', exists and has first derivative as well. Complying with these, problem in (7) becomes a nonlinear convex optimization problem.

- [Conditions for minimaxity] When $\rho(x) = -\log f(x)$, according to Lemma 3.2, the minimum asymptotic variance is achieved. If density f also minimize the objective in (13), the minimum variance is achieved in the worst scenario. Such an estimate is called an asymptotic minimax estimate. From Lemma 3.3, the above mentioned minimizer f should satisfy equation (15).

Readers can verify that the following function is a solution to equation (15).

$$f_0(x) \quad = \quad \begin{cases} c\left[\cos\lambda_1\frac{x}{\delta}\right]^2, & |x| < \delta, \\ c\cdot\exp\left(-2\lambda_2\frac{|x|}{\delta}\right)\cdot\cos^2\lambda_1\cdot\exp(2\lambda_2), & |x| \geq \delta, \end{cases} \tag{16}$$

where $0 < \lambda_1 < \frac{\pi}{2}$, $\lambda_2 > 0$. The above is constructed by considering the general solutions to the differential equation (15). One of the simplest form that satisfies all the aforementioned conditions is chosen. Special care is given to ensure that $\log(f_0)$ has second derivative, as readers will see later. More discussion regarding our choice of function f_0, especially how it differs from Huber's estimator, will be provided in Section 5.7.

Recall $\rho = -\log f_0$, we have

$$\rho(x) = \begin{cases} -\log c - 2\log\cos\lambda_1\frac{x}{\delta}, & |x| < \delta, \\ -\log c + 2\lambda_2\frac{|x|}{\delta} - 2\lambda_2 - 2\log\cos\lambda_1, & |x| \geq \delta. \end{cases} \tag{17}$$

Note $\rho(x)$ can be simplified without changing the optimization problem in (7): i.e., replacing $\rho(x)$ with $a\rho(x) + b, a > 0$ in (7) gives an equivalent optimization problem. Note that $\rho(x)$ is linear outside the interval $[-\delta, \delta]$.

5.3.5 Parameters in RSBN

The parameters $c, \delta, \alpha, \lambda_1, \lambda_2$ satisfy the following conditions:

$$\begin{aligned} \int_{-\delta}^{\delta} f_0(x)dx &= 1-\alpha; \\ \lim_{x\to\delta+} f'(x) &= \lim_{x\to\delta-} f'(x); \end{aligned} \tag{18}$$

or equivalently,

$$\lim_{x\to\delta+} \rho'(x) = \lim_{x\to\delta-} \rho'(x); \tag{19}$$

$$\int_{\delta}^{+\infty} f_0(x)dx = \frac{\alpha}{2}. \tag{20}$$

From (19), we have

$$\lambda_2 = \lambda_1 \tan\lambda_1. \tag{21}$$

From (18) and (20), we have

$$\begin{aligned} 1-\alpha &= c\int_{-\delta}^{\delta}\left[\cos\lambda_1\frac{x}{\delta}\right]^2 \\ &= \frac{c\delta}{2}\left(\frac{1}{\lambda_1}\sin 2\lambda_1 + 2\right); \qquad (22) \\ \frac{\alpha}{2} &= c\cdot\cos^2\lambda_1\cdot\exp(2\lambda_2)\int_{\delta}^{+\infty}\exp\left(-2\lambda_2\frac{x}{\delta}\right) \\ &= \frac{c\delta}{2}[\cos\lambda_1]^2\frac{1}{\lambda_2}, \qquad (23) \end{aligned}$$

respectively. From (22), (23) and (21), we have

$$\begin{aligned} \frac{\alpha}{1-\alpha} &\overset{(22),23)}{=} \frac{\frac{1}{\lambda_2}\cdot\cos^2\lambda_1}{1+\frac{1}{2\lambda_1}\sin 2\lambda_1} \\ &\overset{(21)}{=} \frac{\frac{1}{\lambda_1}\cos^3\lambda_1/\sin\lambda_1}{1+\frac{1}{2\lambda_1}\sin 2\lambda_1} \\ &= \frac{\cos^3\lambda_1}{\lambda_1\cdot\sin\lambda_1+\sin^2\lambda_1\cdot\cos\lambda_1}. \end{aligned}$$

Hence

$$\alpha = \frac{\cos^3\lambda_1}{\lambda_1\cdot\sin\lambda_1+\cos\lambda_1}. \tag{24}$$

Proposition 3.4 *The proportion α defined in the stochastically bounded noises condition and the parameter λ_1 in RSBN have a relation stated in (24).*

Figure 48 illustrates the relationship between α and λ_1.

Now we consider a simplified version of (17). As an objective function in (7), the following ρ is equivalent to the one in (17).

$$\rho(x) = \begin{cases} -\log\cos\lambda_1\frac{x}{\delta}, & |x|<\delta; \\ \lambda_2\frac{|x|}{\delta}-\lambda_2-\log\cos\lambda_1, & |x|\geq\delta. \end{cases} \tag{25}$$

Bringing in (21), we get exactly the expression in (8). Up to this point, we have established the Theorem 2.2.

We summarize the procedure of getting function ρ for RSBN. By some prior information, we know the values of α and δ. From (24), we can compute for λ_1. From (21), we can compute for λ_2. Substituting values λ_1 and λ_2 into (25), we have the close form formula for ρ. The following flow chart summarizes how to get ρ from α:

$$\alpha,\delta \xrightarrow{(24)} \lambda_1 \xrightarrow{(21)} \lambda_2 \xrightarrow{(25)} \rho.$$

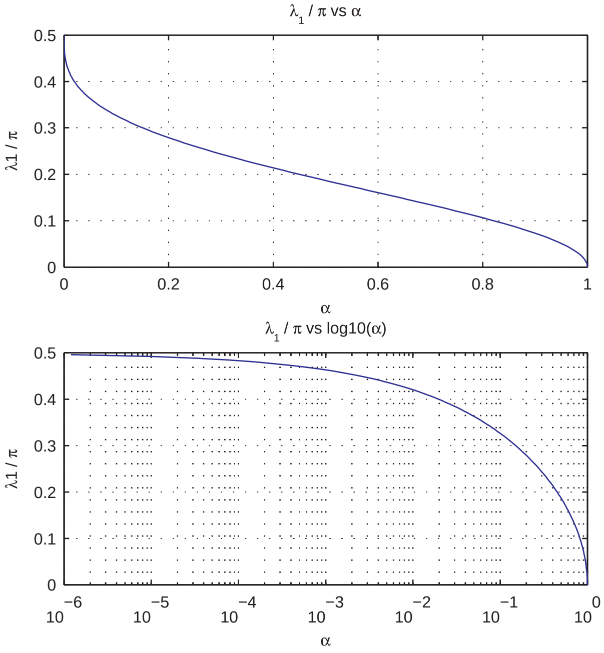


Figure 48: Parameter λ_1 vs. α. The upper one is ordinary; the bottom takes $\log 10$ on α.

5.3.6 Fisher Information of the Least Informative Distribution

We consider two important quantities associated with RSBN: Fisher information and asymptotic variance. For Fisher information, we give a close form solution with respect to λ_1. Since we know the relationship between λ_1 and α in (24), we have the relationship between the Fisher information and α. Figure 49 will illustrate it. For the asymptotic variance, we need to know the exact noise distribution. In the next subsection, we describe how to compute it in a general case.

We start with the Fisher information $I(f_0)$. We consider the location estimation case. Let $f_\theta = f_0(x - \theta)$, where f_0 is the least informative distribution in Section 5.3.4. We have

$$I(f_0) = 4\frac{\lambda_1^2}{\delta^2}\frac{\lambda_1 \cdot \sin\lambda_1}{\lambda_1 \cdot \sin\lambda_1 + \cos\lambda_1}. \tag{26}$$

The details in validating the above equation is postponed to Appendix A.1. Taking $\delta = 1.0$ and combining (24) and (26), we have the relationship between the Fisher information $I(f_0)$ and α. Since $\lambda_1 \in [0, \frac{\pi}{2}]$, the range of Fisher information $I(f_0)$, based on (26), is from 0 to π^2/δ^2. Figure 49 shows the relationship between α and the Fisher information $I(f_0)$. It is easy to find that small α leads to large Fisher information.

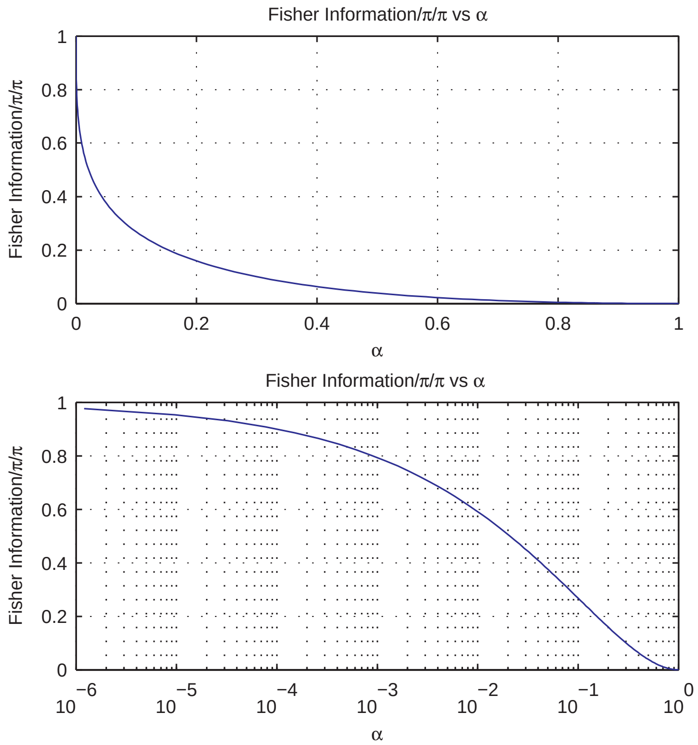


Figure 49: Fisher information $I(f_0)$ versus α. The upper one takes ordinary coordinates; the lower takes $\log_{10}$ on α.

5.3.7 Asymptotic Variance of RSBN

As for asymptotic variance, in Section 5.3.1 we have already known that

$$\text{asymptotic variance} = \frac{\int \psi^2 dF}{(\int \psi' dF)^2}. \tag{27}$$

Since $\psi = \rho'$, we have

$$\psi(x) = \rho'(x) \overset{(25)}{=} \begin{cases} \frac{\lambda_1}{\delta} \tan \lambda_1 \frac{x}{\delta}, & |x| < \delta; \\ \operatorname{sign}(x) \cdot \frac{\lambda_1}{\delta} \tan \lambda_1, & |x| \geq \delta; \end{cases} \tag{28}$$

and

$$\psi'(x) = \begin{cases} \frac{\lambda_1^2}{\delta^2} \sec \lambda_1 \frac{x}{\delta}, & |x| < \delta; \\ 0, & |x| \geq \delta. \end{cases}$$

Note ψ' is no longer continuous. As long as the noise has probability density function f that makes (27) meaningful, we can compute the asymptotic variance of the RSBN estimate.

5.3.8 Robustness

We now consider the robust property of RSBN. We compute the *breakdown* point—the maximum proportion of observations that can be arbitrarily distorted, while the estimate still does not "blow up" (i.e., not going to ±∞).

On page 16 in [43], we know that

$$\text{breakdown point} = \epsilon^\star = \frac{\eta}{1+\eta},$$

where $\eta = \min\left\{-\frac{\psi(-\infty)}{\psi(+\infty)}, -\frac{\psi(+\infty)}{\psi(-\infty)}\right\}$. From the formula of ψ in the last section, we have $\eta = 1$. Hence $\epsilon^\star = 1/2$, which is the largest breakdown point we can have for M-estimates.

Lemma 3.5 *The breakdown point of the RSBN estimate is* 1/2*.*

5.4 Numerical Algorithm: Proximal Point Method

In this subsection, we describe a proximal point algorithm. The purpose is to give readers who may not have access to a sophisticated optimization software package an extremely–easy–to–use algorithm.

The rest of this section is organized as follows. Section 5.4.1 describes the general idea of a proximal point method. The RSBN can be formulated as a *partial inverse problem*, which is described in Section 5.4.2. Section 5.4.3 describes how to solve a partial inverse problem. An algorithm that solves RSBN is provided in Section 5.4.4. Some analysis regarding the convergence rate of the proposed algorithm is presented in Section 5.4.5.

5.4.1 General Idea

The proximal point algorithm solves the following problem:

$$\text{Find } \mu \in \mathbf{R}^n \quad : \quad 0 = \mathbf{U}(\mu), \tag{29}$$

$$\text{where} \qquad \mathbf{U} : \mathbf{R}^n \to \mathbf{R}^n \text{ is an operator.}$$

The proximal point algorithm includes two steps:

```
Algorithm to Solve $0 = \mathbf{U}(\mu)$.

1. Choose $\mu^{(0)}, n = 0$.

2. Repeat

       $\mu^{(n+1)} = (\mathbf{I} + \mathbf{U})^{-1}\mu^{(n)}$,
       $n = n + 1$,

   Until convergence.
```

Here $\mathbf{I}$ is the identity operator, and $(\mathbf{I} + \mathbf{U})^{-1}$ is the inverse of operator $(\mathbf{I} + \mathbf{U})$. The following results are known [89].

- Let μ^0 denote the solution to (29), i.e., $0 = \mathbf{U}(\mu^0)$. If $\{\mu^{(n)}\}$ converges, then it converges to μ^0.
- If $\mathbf{U}$ is a monotone operator in $\mathbf{R}^n$, then $(\mathbf{I} + \mathbf{U})^{-1}$ is well defined. (Operator $\mathbf{U}$ is a monotone operator if for any $x_1, x_2 \in \mathbf{R}^n$, the inner product $\langle x_1 - x_2, \mathbf{U}(x_1) - \mathbf{U}(x_2)\rangle \geq 0$.)
- If $\mathbf{U}$ is a monotone operator in $\mathbf{R}^n$, then $\{\mu^{(n)}\}$ converges.

5.4.2 Partial Inverse

Problem (7) can be cast as a *partial inverse problem*. Suppose A is a subspace of $\mathbf{R}^n$, $A \subset \mathbf{R}^n$ and B is the perpendicular compliment of A, $B = A^{\perp}$. The *partial inverse problem* is:

$$\text{find } x, y \in \mathbf{R}^n : \begin{cases} x \in A, \\ y \in B, \\ y = \mathbf{U}(x). \end{cases} \tag{30}$$

If $\mathbf{U}$ is strictly monotone, the solution of the *partial inverse* is unique.

Problem (30) can be formulated as (29). Suppose $x, y \in \mathbf{R}^n$ have decomposition:

$$x = x_A + x_B, \qquad y = y_A + y_B,$$

where $x_A, y_A \in A$ and $x_B, y_B \in B$. We define a new operator $\mathbf{U}_A$, such that $x_B + y_A = \mathbf{U}_A(x_A + y_B)$ if and only if $y = y_A + y_B = \mathbf{U}(x_A + x_B) = \mathbf{U}(x)$. Suppose z has a decomposition: $z = z_A + z_B$, where $z_A \in A, z_B \in B$. A general theorem says that (x, y) is the solution to (30) if and only if

$$\exists z : 0 = \mathbf{U}_A(z), \tag{31}$$

where $x = z_A, y = z_B$. By solving (31), we get an exact solution to (30). Note that (31) has the same form as (29).

5.4.3 Solving Partial Inverse

Based on (31) and the algorithm in Section 5.4.1, the key to solving a *partial inverse problem* is to find $(\mathbf{I}+\mathbf{U}_A)^{-1}$. Following the notations in Section 5.4.2, since $x_B+y_A = \mathbf{U}_A(x_A+y_B)$, we have $x+y = (\mathbf{I}+\mathbf{U}_A)(x_A+y_B)$. In other words, $x_A+y_B = (\mathbf{I}+\mathbf{U}_A)^{-1}(x+y)$. In order to solve $(\mathbf{I}+\mathbf{U}_A)^{-1}(u)$, if we can find (x, y) satisfying

$$\begin{cases} u = x + y, \\ y = \mathbf{U}(x), \end{cases}$$

then $(\mathbf{I}+\mathbf{U}_A)^{-1}(u) = x_A + y_B$. Since $u = x + y = (\mathbf{I}+\mathbf{U})(x)$, we have

$$\begin{cases} x = (\mathbf{I}+\mathbf{U})^{-1}(u), \\ y = u - x. \end{cases}$$

Now we have the algorithm to solve $(\mathbf{I}+\mathbf{U}_A)^{-1}$.

Algorithm to Solve $(\mathbf{I}+\mathbf{U}_A)^{-1}$.

- Find x, so that $x = (\mathbf{I}+\mathbf{U})^{-1}(u)$;
- Let $y = u - x$;
- $(\mathbf{I}+\mathbf{U}_A)^{-1}(u) = x_A + y_B$.

Note this is a general method to solve (30). If $(\mathbf{I}+\mathbf{U})^{-1}$ is easy to implement, then $(\mathbf{I}+\mathbf{U}_A)^{-1}$ is easy to implement.

5.4.4 Application to RSBN

Now we apply the previously developed method to RSBN. Consider the first order necessary condition of (7), we have

$$0 = A^T \psi(Ax - y), \tag{32}$$

where $\psi = \rho'$, $\psi(y - Ax) = [\psi((y - Ax)_1), \psi((y - Ax)_2), \cdots, \psi((y - Ax)_n)]^T$, $(y - Ax)_i$ denotes the ith component of vector $y - Ax$, and ψ is defined in (28). Equation (32) is equivalent to

$$\text{find } u, v : \begin{cases} u = Ax, \\ v = \psi(u - y), \\ 0 = A^T v. \end{cases} \tag{33}$$

In other words,

$$\text{find } u, v : \begin{cases} u \in \text{Range}(A), \\ v \in \text{Kernel}(A), \\ v = \psi(u - y). \end{cases}$$

Following the algorithm in Section 5.4.3, we have

Algorithm for RSBN

1. **Choose** $\mu^{(0)} \in \mathbf{R}^n, k = 0$.
2. **Find** u_i, such that $\psi(u_i - y_i) + u_i = \mu_i^{(k)}, i = 1, 2, \cdots, n$.
3. **Let** $v_i = \mu_i^{(k)} - u_i, i = 1, 2, \cdots, n$.
4. **Project** $u = (u_1, \cdots, u_n)^T, v = (v_1, \cdots, v_n)^T$.

   $$\begin{aligned} \mu^{(k+1)} &= \mathbf{P}_A(u) + \mathbf{P}_{Ker(A)}(v) \\ &= v + A(A^TA)^{-1}A^T(u - v). \end{aligned}$$

   Here $\mathbf{P}_A$ and $\mathbf{P}_{Ker(A)}$ are projection operators to subspaces range of A and kernel of A respectively.
5. If not converge, $k = k + 1$, go back to step 2.

In step 2, since ψ in (28) is monotone increasing, x_i will have a unique solution. But because there is a tangent function in ψ in RSBN, one needs to implement a line search algorithm to solve it. We can see that if function ψ is piecewise polynomial, this method is quite appealing, because a close form solution is available to the equation in step 2. This approach has been used in solving Huber's M-estimate, see [67].

After getting u, the x can be solved via $u = Ax$. Recall matrix A is of full column rank.

5.4.5 Analysis

It is possible that the above mentioned algorithm converges slowly to the solution. Here is an example. Assume the projection matrix associated with $\mathbf{P}_A$ is, for $d < n$,

$$\begin{pmatrix} \begin{pmatrix} 1 & & \\ & \ddots & \\ & & 1 \end{pmatrix}_{d\times d} & 0 \\ 0 & 0 \end{pmatrix}_{n\times n};$$

i.e., it projects to the first d coordinates. Following the notations in the above subsection, we have

$$\mu_i^{(k+1)} = \begin{cases} u_i, & 1 \leq i \leq d, \\ v_i, & 1 + d \leq i \leq n. \end{cases}$$

Restricted to $1 \leq i \leq d$, we have

$$|\mu_i^{(k+1)} - \mu_i^{(k)}| = |u_i - \mu_i^{(k)}| = |\psi(u_i - y_i)| \leq \frac{\lambda_1}{\delta} \tan \lambda_1,$$

where the last term is a constant, the second equality is based on the step (b) in the RSBN algorithm, and the inequality is based on (28). If u^0 is the solution of the RSBN, assuming that we started with an all zero vector $\mu^{(0)} = (0, 0, \ldots, 0)^T$, the proposed proximal point algorithm takes at least

$$\frac{\max_{1 \leq i \leq d} |u_i^0|}{\lambda_1/\delta \tan \lambda_1}$$

steps to converge. Note the number of steps can be large, if the maximum entry $\max_{1 \leq i \leq d} |u_i^0|$ is large.

The reason that the proximal point approach can be slow is that it does not take advantage of high degree smoothness of the objective function. For example, it does not use the second derivatives. More efficient numerical solution can be developed by taking advantage of the existence of second derivatives. Most of state-of-the-art optimization software will do so automatically. We propose one alternative in the next section.

5.5 Other Implementation: SNOPT and SQP

As an alternative, we use some state-of-the-art optimization software to solve the RSBN *directly.* In this research, we use a general-purpose optimization package—SNOPT. It is a software package developed in [33]. It minimizes a linear or nonlinear function subject to bounds on the variables, as well as sparse linear or nonlinear constraints. It is suitable for large-scale linear and quadratic programming and for linearly constrained optimization, as well as for general nonlinear programs. In our case, in (7), we have linear constraints and a nonlinear but convex objective function.

SNOPT finds a solution that is *locally optimal.* Ideally, any nonlinear functions should be smooth and users should provide gradients. In our case, since the objective function in

(7) is convex, the *locally optimal* solution will coincide with the global optimal solution. For RSBN, the gradients are given in (28).

SNOPT uses a sequential quadratic programming (SQP) algorithm that obtains a search direction from a sequence of quadratic programming subproblems. Each QP subproblem minimizes a quadratic model of a certain Lagrangian function subject to a linearization of the constraints. An augmented Lagrangian merit function is reduced along each search direction to ensure convergence from any starting point.

The source code for SNOPT is written in Fortran. In order to use it, a Fortran compiler is required. The numerical examples in the present chapter are a result of combining some MATLAB programming, Unix shell programming, Fortran programming, and SNOPT.

5.6 Simulation

5.6.1 An Illustrative Example: Variable Star

In this section, we study a well-known data set in the time series analysis — magnitudes of a variable star at midnight on 600 successive nights. [7] showed that it is a superposition of two 'dominant' sinusoid functions. We are taking a slightly different viewpoint. We assume that the underlying signal (denoted by $\mathbf{s}$) is a smooth signal residing in a low dimensional linear subspace. The observed magnitudes denoted by $\mathbf{y}$, $\mathbf{y} = (y_1, y_2, \ldots, y_{600})^T$, are an approximation to s. In our case, $\mathbf{y}$ is the rounded version of s: i.e., $y_i = [s_i + 0.5]$, where $[x]$ is the largest integer no larger than x. It is evident that the mapping from $\mathbf{s}$ to $\mathbf{y}$ is completely nonlinear. Let $\mathbf{y} = \mathbf{s}+\mathbf{n}$, where $\mathbf{n}$ is the so-called noise sequence. Considering the source where the noise sequence is generated, the Gaussian assumption on the distribution of $\mathbf{n}$ is not appropriate. We assume that the subspace, on which the signal resides, is known to us. In this case, we compare ordinary least square estimate, Huber's estimate, and RSBN.

We consider the discrete cosine transform (DCT). The DCT with signal length n has the kth basis function:

$$c_k(i) = \begin{cases} \sqrt{1/n}, & k = 0, i = 1, 2, \cdots, n; \\ \sqrt{2/n} \cos[(i - \frac{1}{2})k\frac{\pi}{n}], & k \neq 0, i = 1, 2, \cdots, n. \end{cases}$$

The reasons of choosing DCT are: (a) DCT is a real analogous of the Fourier transform, which is widely adopted in representing cyclic signals; (b) there are fast numerical algorithms to implement DCT.

First, we study the original variable star data set. We find the subspace that contains most of the signal's energy. This can be done by carrying out a DCT transform, then retaining the coefficients with the largest amplitudes. Later, we intentionally distort the observation. Three different ways of projection are then compared. We illustrate the optimality of our method.

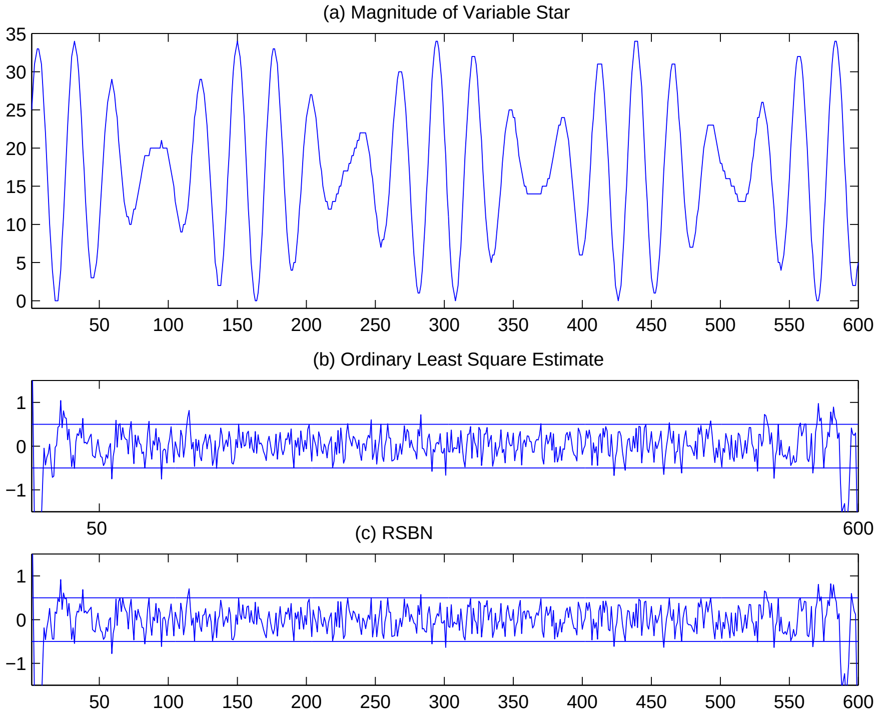


Figure 50: From top to bottom: (a) Integer-valued magnitude of a variable star at midnight on 600 successive nights; (b) The deviation (between estimation and observation) corresponding to the ordinary least square estimate; And (c) the deviation corresponding to RSBN.

Figure 50 (a) shows the magnitude vector $\mathbf{y}$ of the variable star. We take a DCT of $\mathbf{y}$, keep the 10% of coefficients that have the 10% largest amplitudes (of coefficients). The associated 10% basis functions span the subspace that contains the largest possible proportion of the energy. We denote the subspace by A. The dimension of A is 60. Projecting the observation $\mathbf{y}$ to A by the ordinary least square regression, we have $\mathbf{P}_{A,LS}(\mathbf{y}) = \hat{s}_{LS,1}$, where subscript ‘LS’ indicates least square estimate, and ‘1’ indicates for original observation $\mathbf{y}$. The deviation between the original sequence $\mathbf{y}$ and the estimate $\hat{s}_{LS,1}$ is illustrated in Figure 50 (b). Then we project the observation $\mathbf{y}$ to A by using RSBN. We choose $\delta = 0.5$, $\lambda_1 = 0.46\pi$ and ρ is given in (25). We denote $\mathbf{P}_{A,RSBN}(\mathbf{y}) = \hat{s}_{RSBN,1}$, where subscript ‘RSBN’ indicates a RSBN estimate. The deviation, $\mathbf{y} - \hat{s}_{RSBN,1}$, is illustrated in Figure 50 (c). Since the deviations are supposed to be round off errors, ideally they should be within the interval $[-0.5, 0.5]$. For the least square estimate in Figure 50 (b), there are 70 deviations having amplitudes larger than 0.5, and 16 of them having amplitudes larger than 1.0. For RSBN in Figure 50 (c), there are 44 deviations having amplitudes larger than 0.5, and 15 of them having amplitude larger than 1.0. In this case, compared to the ordinary least square estimate, the RSBN has less deviations falling beyond the ideal interval $[-0.5, 0.5]$. Of course, at the same time, we should observe a loss in the mean square error, which is what a least square approach tries to minimize. The sum of squares of deviations in the least square estimate is 10.4337, and the one for RSBN is 10.6022.

Now we randomly pick up two positions in the variable star sequence. In particular, we choose position 224 and 446. Originally, $y_{224} = 15$ and $y_{446} = 16$. Suppose the decimal points in these numbers were somehow misspecified. The recorded values become $y'_{224} = 1.5$ and $y'_{446} = 160$. Without loss of generality, let $\mathbf{y}'$ denote the new sequence. Figure 51 shows $\mathbf{y}'$.

Recall $\mathbf{P}_{A,LS}$ and $\mathbf{P}_{A,RSBN}$ denote the projection operators to subspace A by least square estimate and RSBN respectively. Let $\mathbf{P}_{A,H}$ denote a projection operator to A via Huber’s M-estimate. Recall that a Huber’s M-estimate is for $\Delta > 0$, choose

$$\rho(x) = \begin{cases} x^2, & |x| < \Delta, \\ 2\Delta|x| - \Delta^2, & |x| \geq \Delta, \end{cases}$$

in (7) [43, 44]. In Huber’s estimate, the function ρ is piecewise linear (outside a neighborhood of the origin) or quadratic (inside a neighborhood of the origin).

Consider the projections $\hat{s}_{LS,2} = \mathbf{P}_{A,LS}(\mathbf{y}')$, $\hat{s}_{RSBN,2} = \mathbf{P}_{A,RSBN}(\mathbf{y}')$, and $\hat{s}_{H,2} = \mathbf{P}_{A,H}(\mathbf{y}')$. The deviations $\mathbf{y} - \hat{s}_{LS,2}$, $\mathbf{y} - \hat{s}_{RSBN,2}$ and $\mathbf{y} - \hat{s}_{H,2}$ are plotted in Figure 52 (a), (b), and (c). Note these are the deviations from the estimates to the “original” signal

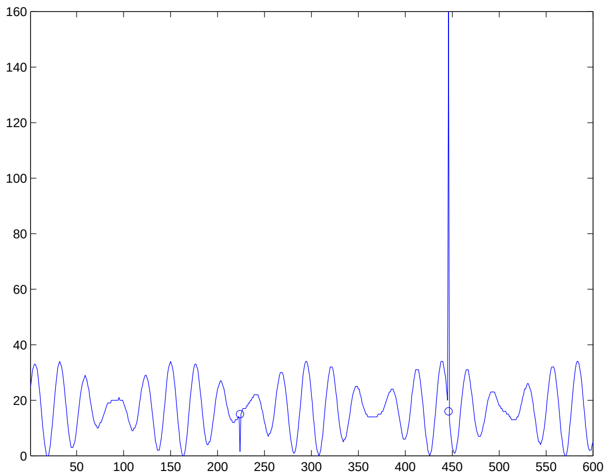


Figure 51: The distorted variable star signal. On day 224 and 446, the decimal points in the observed values were wrongly shifted to the left (15 → 1.5) and right (16 → 160) respectively. Circles indicate the true values.

Table 1: Some statistics for three regression methods for the distorted variable star data.

	ordinary least square	Huber's M-estimate	*RSBN*
Square root of sum of squares, $\|\mathbf{y} - \hat{s}_{*,2}\|_2$	47.2771	10.6658	10.6053
Number of amplitudes > 0.5	372	53	45
Number of amplitudes > 1.0	196	16	15

sequence $\mathbf{y}$ (not $\mathbf{y}'$). Table 1 shows some statistics on the performance of three different methods.

There are several phenomena noteworthy. First of all, the deviation of the least square estimate is significantly worse than the other two. This illustrates that least square estimate is not a robust method. Second, the performance of RSBN has almost no difference between the two cases: $\mathbf{y}$ and $\mathbf{y}'$. In other words, $\hat{s}_{RSBN,2}$ is as close to $\mathbf{y}$ as $\hat{s}_{RSBN,1}$ is. Third, RSBN performs nearly as well as the Huber's M-estimate. RSBN is slightly better. It is not surprising that the performances of RSBN and Huber's M-estimate are close, because the objective functions in (7) for these two are very close to each other. One commonality: they both take linear function outside an interval: $(-\delta, \delta)$ for RSBN and $(-\Delta, \Delta)$ for Huber's.

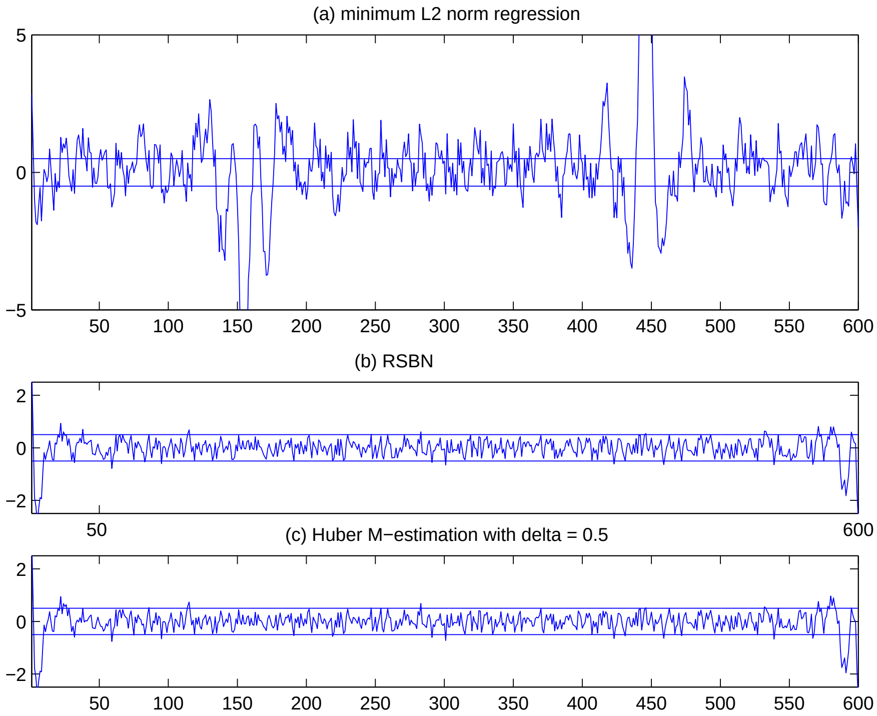


Figure 52: Differences between the original variable star signal and the estimates from distorted signal by three different methods. The corresponding methods are, from top to bottom: (a) ordinary least square regression, (b) RSBN, and (c) Huber's M-estimate with $\Delta = 0.5$.

5.6.2 Comparison with Ordinary Least Square Estimate and Huber's Estimate

We compare three different regression methods: ordinary least square estimate, RSBN, and Huber's M-estimate. We demonstrate that for distorted Gaussian noises, RSBN does the best job.

Recall we have a linear model:

$$\mathbf{y} = A\mathbf{x} + \varepsilon, \tag{34}$$

where $A \in \mathbf{R}^{n\times m}$ is the model matrix, $\mathbf{x} \in \mathbf{R}^m$ is the parameter vector, $\varepsilon \in \mathbf{R}^n$ is the noise vector, and $\mathbf{y} \in \mathbf{R}^n$ is the observation vector. In this experiment, we choose $m = 15, n = 600$.

5.6.2.1 Design of Simulation

In each experiment, for the model in (34), A is generated by sampling each entry ($A_{ij}, 1 \leq i \leq n, 1 \leq j \leq m$) from a standard Normal distribution (Normal$(0, 1)$), with the constraint that matrix A must have full column rank. If the generated matrix A does not have full column rank, the process is repeated instead of proceeding to the next step. The vector $\mathbf{x}$ is generated as a standard Normal vector in $\mathbf{R}^m$, $\mathbf{x} \sim$ Normal$(0, \mathbf{I}_m)$. The vector ε is generated as a standard Normal vector in $\mathbf{R}^n$, $\varepsilon \sim$ Normal$(0, \mathbf{I}_n)$. The observation vector $\mathbf{y}$ is a superposition: $\mathbf{y} = A\mathbf{x} + \varepsilon$.

Let span(A) denote the linear subspace spanned by the column vectors in matrix A. Obviously, it has m degrees of freedom, dim(span(A)) $= m$. Recall the operator $\mathbf{P}_{A,LS} : \mathbf{R}^n \to$ span(A) is the projection operator from Euclidean space $\mathbf{R}^n$ to the linear subspace span(A). In other words,

$$\mathbf{P}_{A,LS}(\mathbf{y}) = \underset{\mathbf{u} \in \text{span}(A)}{\text{argmin}} \quad \|\mathbf{u} - \mathbf{y}\|_{\ell_2}^2.$$

Let $d_{LS,1}$ denote the deviation vector from the least square projection $\mathbf{P}_{A,LS}(\mathbf{y})$ to the true linear component $A\mathbf{x}$. Note here the first subscript "LS" indicates the least square method, and the second subscript "1" indicates the Gaussian noise vector (ε). We have

$$d_{LS,1} = \mathbf{P}_{A,LS}(\mathbf{y}) - A\mathbf{x} = \mathbf{P}_{A,LS}(\varepsilon). \tag{35}$$

We then distort the Gaussian vector ε. We randomly select 1% entries in ε, multiply them by 200 (value 200 is arbitrarily chosen). The new vector is denoted by ε'. Effectively, each entry of ε' follows a mixed normal distribution: $\varepsilon'_i \sim 0.99$Normal$(0, 1) + 0.01$Normal$(0, 200^2), 1 \leq i \leq n$. Denote $\mathbf{y}' = A\mathbf{x} + \varepsilon'$.

Recall previously mentioned notations, $\mathbf{P}_{A,RSBN} : \mathbf{R}^n \to$ span(A) and $\mathbf{P}_{A,H} : \mathbf{R}^n \to$ span(A) are projection operators by adopting RSBN and Huber's M-estimate respectively. Let $d_{LS,2}$, $d_{RSBN,2}$ and $d_{Huber,2}$ denote the deviation vectors corresponding to the least square estimate, RSBN, and Huber's M-estimate, respectively. (The first subscripts of the above d's indicate methods, and the second subscript "2" indicates distorted noise vector

ε'.) We have

$$\begin{aligned} d_{LS,2} = &\quad \mathbf{P}_{A,LS}(\mathbf{y}') - A\mathbf{x} &= \mathbf{P}_{A,LS}(\varepsilon'); \\ d_{RSBN,2} = &\quad \mathbf{P}_{A,RSBN}(\mathbf{y}') - A\mathbf{x} &= \mathbf{P}_{A,RSBN}(\varepsilon'); \\ d_{Huber,2} = &\quad \mathbf{P}_{A,Huber}(\mathbf{y}') - A\mathbf{x} &= \mathbf{P}_{A,Huber}(\varepsilon'); \end{aligned}$$

We repeat the experiments for 1000 times. Each time, for the distorted noises, three methods lead to three deviation vectors: $d_{LS,2}, d_{RSBN,2}$ and $d_{Huber,2}$. Let $d_{LS,2}^{(i)}, d_{RSBN,2}^{(i)}$ and $d_{Huber,2}^{(i)}$ denote the deviation vectors we get in the ith experiment, we have totally 3000 n-D vectors: $d_{LS,2}^{(i)}, d_{RSBN,2}^{(i)}, d_{Huber,2}^{(i)}, i = 1, 2, \cdots, 1000$.

The smaller the deviations are, the better the regression method is. In the multivariate situation, we need to quantify the smallness. We will report our comparison in Section 5.6.2.3.

5.6.2.2 Cut-off Value

To measure the robustness of different methods, it is nature to compare the deviation vectors $d_{LS,2}, d_{RSBN,2}$, and $d_{Huber,2}$ with deviation vector $d_{LS,1}$, because $d_{LS,1}$ is the deviation of an ideal method (least square estimation, or MLE) in the ideal situation (with Gaussian noises). We propose to study the number of deviations with amplitudes above a quantity τ: i.e., for $1 \le i \le 1000$,

$$\#\{j : \quad |(d_{*,2}^{(i)})_j| \quad > \tau, 1 \le j \le n\},$$

where $*$ can be LS, RSBN, or Huber. Here notation # stands for the cardinality of a finite set. The jth component of vector $d_{*,2}^{(i)}$ is denoted as $(d_{*,2}^{(i)})_j$. Value τ can be viewed as a quantile of random variable $\|d_{LS,1}\|_\infty$. The value τ will be called a *cut-off* value.

The following is to derive a reasonable value of τ. We study the distribution of random variable $\|d_{LS,1}\|_\infty$. Let $\|d_{LS,1}\|_2$ denote the ℓ_2 norm of the vector $d_{LS,1}$. We have

$$\|d_{LS,1}\|_\infty = \|d_{LS,1}\|_2 \cdot \frac{\|d_{LS,1}\|_\infty}{\|d_{LS,1}\|_2}.$$

We list three facts. For details, please refer to Appendix A.2.

- Random variables $\|d_{LS,1}\|_2$ and $\|d_{LS,1}\|_\infty / \|d_{LS,1}\|_2$ are independent;

- Random variable $\|d_{LS,1}\|_2^2$ satisfies the χ_m^2 distribution with m degrees of freedom. Recall m is the column rank of matrix A.

- Assume the projection $\mathbf{P}_{A,LS}$ related to model matrix A has eigenvalue decomposition

$$\mathbf{P}_{A,LS} = U^T \begin{pmatrix} \mathbf{I}_m & \\ & 0 \end{pmatrix} U,$$

  where matrix U is orthogonal. Let

$$\mathbf{x} = U^T \begin{pmatrix} x_m \\ 0_{(n-m)\times 1} \end{pmatrix},$$

  where vector x_m is Uniform on the unit sphere in $\mathbf{R}^m$, $\|x_m\|_2 = 1$, and vector $0_{(n-m)\times 1}$ is an all zero vector. Let $\rho_{\max,m} = \|\mathbf{x}\|_\infty$. The ratio $\|d_{LS,1}\|_\infty / \|d_{LS,1}\|_2$ has the same distribution as $\rho_{\max,m}$. The analytical solution to the probability density function of $\rho_{\max,m}$ could be too complicated to be useful though.

Based on the above three facts, we can find the distribution of $\|d_{LS,1}\|_\infty$ and the cut-off value through simulations. In this chapter, we choose the cut-off value: $\tau = 1$. The related probability $P\{\|d_{LS,1}\|_\infty > \tau\}$ is approximately 3.1×10^{-4}, which is obtained through $100,000$ times of simulations.

5.6.2.3 Simulation Results

Figure 53 illustrates the results from all the steps of one simulation.

- Figure 53 (a) shows the Gaussian noise vector ε. Each element of it satisfies distribution Normal$(0, 1)$.

- Figure 53 (b) shows the deviation vector ($d_{LS,1}$) of the least square regression in the Gaussian noise (ε) case.

- Figure 53 (c) shows the distorted Gaussian noise vector ε'. The vector ε' is gotten by multiplying randomly picked six elements of vector ε by 200.

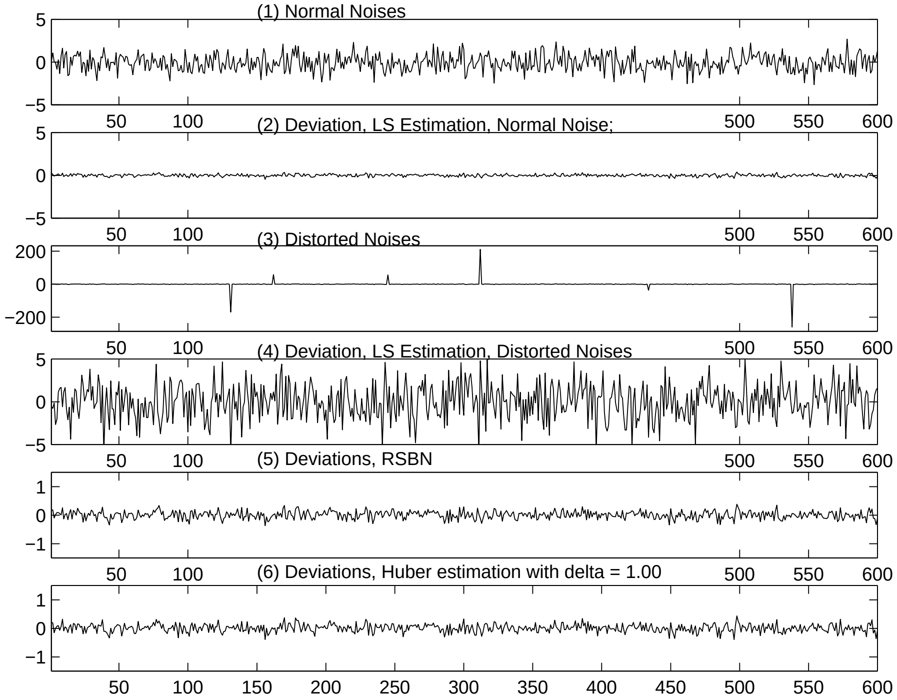


Figure 53: (a) Standard Gaussian noise vector ε, (b) the deviation vector $d_{LS,1}$ of least square regression in the Gaussian noise ε case, (c) the distorted Gaussian noise vector ε', (d) the deviation vector $d_{LS,2}$ from the least square regression with the distorted Gaussian noise ε', (e) the corresponding deviation vector $d_{RSBN,2}$ from the RSBN, (f) the corresponding deviation vector $d_{Huber,2}$ from the Huber's estimate.

- Figure 53 (d) shows the deviation vector $d_{LS,2}$ from the least square regression with the distorted Gaussian noise ε'.
- Figure 53 (e) shows the corresponding deviation vector $d_{RSBN,2}$ from the RSBN.
- Figure 53 (f) shows the corresponding deviation vector $d_{Huber,2}$ from the Huber's estimate.

Continued from Section 5.6.2.1, we get 3000 deviation vectors out of 1000 simulations: $d^{(i)}_{LS,2}$, $d^{(i)}_{RSBN,2}$, $d^{(i)}_{Huber,2}$, $i = 1, 2, \cdots, 1000$.

We choose two ways to compare the three different methods. One is to study the relative ratio of the l_2 norms of a pair of deviation vectors. The other is to count the number of amplitudes above the cut-off line (determined by the τ value developed in Section 5.6.2.2) in each deviation vector.

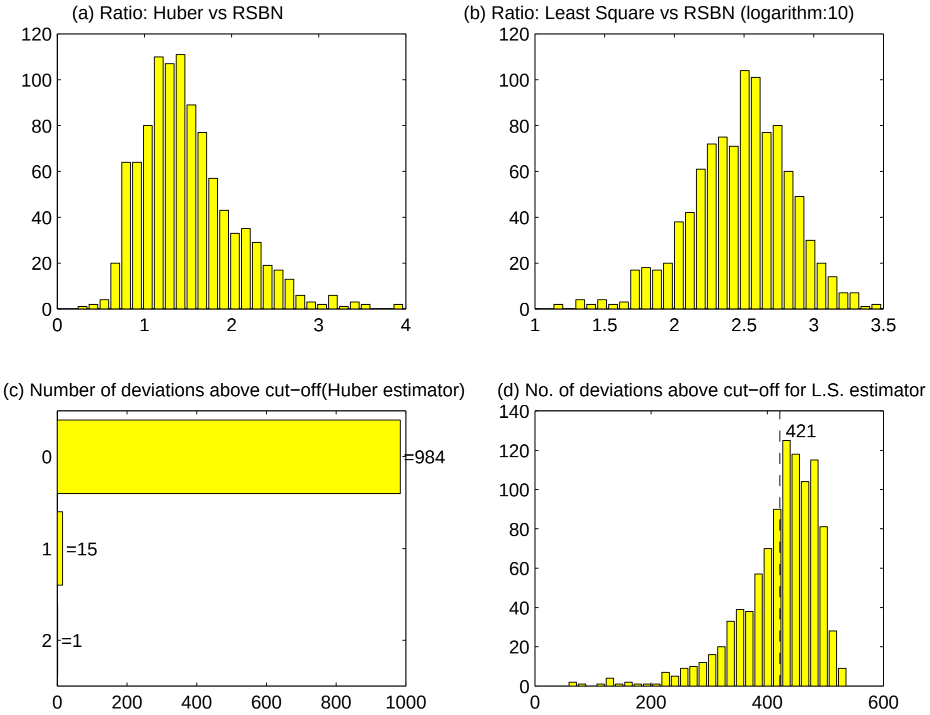


Figure 54: (a) The histogram of the ratios between the Huber's estimate and RSBN: $\|d^{(i)}_{Huber,2}\|_2^2/\|d^{(i)}_{RSBN,2}\|_2^2, i = 1, 2, \cdots, 1000$; (b) The histogram of the logarithm ratios $\log_{10}\left(\|d^{(i)}_{L_2,2}\|_2^2/\|d^{(i)}_{RSBN,2}\|_2^2\right), i = 1, 2, \cdots, 1000$, for the least square regression and the RSBN; (c) For Huber's estimate, number of deviations whose amplitudes are above the cut-off; (d) For the ordinary least square regression, the histogram of number of deviations whose amplitudes are above cut-off.

Figure 54 (a) gives a histogram of the ratios of the ℓ_2 norms of deviation vectors from the Huber's estimate and RSBN: $\|d^{(i)}_{Huber,2}\|_2^2/\|d^{(i)}_{RSBN,2}\|_2^2, i = 1, 2, \cdots, 1000$. We observe that most of them are above 1. This implies the RSBN tends to give smaller sum square of deviations than the Huber's estimate does . Figure 54 (b) shows a histogram of logarithm (base 10) of ratios corresponding to the least square estimate and the RSBN: $\log_{10}\left(\|d^{(i)}_{LS,2}\|_2^2/\|d^{(i)}_{RSBN,2}\|_2^2\right), i = 1, 2, \cdots, 1000$. The reason to take logarithm is that some ratios can be extremely large. Obviously, the least square regression for non-Gaussian noise leads to much higher sum of squares of deviations than the RSBN does.

Define the numbers of amplitudes above the cut-off in the following way:

$$\Gamma^{(i)}_{*,2} = \#\{j : \quad |(d^{(i)}_{*,2})_j| > 1, \quad 1 \leq j \leq n\},$$

where the $*$ can be subscripts: LS, RSBN, or Huber. We observe that for all $1 \leq i \leq 1000$, $\Gamma^{(i)}_{RSBN,2} = 0$. This means the RSBN is very robust (in the sense that there is no outstanding deviation from the true signal). The Huber's estimate performs comparably. Figure 54 (c) gives a histogram of $\Gamma^{(i)}_{Huber,2}, i = 1, 2, \cdots, 1000$. We observe that 15 out of 1000 of them have 1 deviation whose amplitude is larger than 1, and only 1 out of 1000 of them have 2 deviations whose amplitudes are greater than 1. Figure 54 (d) shows a histogram of $\Gamma^{(i)}_{LS,2}, i = 1, 2, \cdots, 1000$. We can see that in most simulations, the number of deviations with amplitudes above the cut-off 1 is large. The average number of deviations with amplitudes above the cut-off is 421, which is roughly 70% of the signal.

In this simulation, the RSBN outperforms the Huber's estimate, and the Huber's estimate outperforms the ordinary least square regression.

5.7 Discussion

5.7.1 A General Regression Formulation

Equation (7) is consistent with many approaches that exist in the literature.

1. If $\rho(x) = x^2$, (7) is the classical least square regression. The solution can be given by applying hat matrix: $\hat{x} = (A^T A)^{-1} A^T y$. We prefer this formulation if the residuals are normally distributed.

2. For $\Delta > 0$, we have

$$\rho(x) = \begin{cases} 0, & |x| < \Delta; \\ |x| - \Delta, & |x| \geq \Delta. \end{cases}$$

   Formulation (7) is an ℓ_1 regression with a 'dead zone'. By adding some slack variables, (7) can be formulated as a linear programming problem. Readers can verify that the following linear programming problem is equivalent to the problem in (7).

$$\begin{array}{rll} \text{minimize} & \sum_{i=1}^{n} t_i, & \\ \text{subject to} & -t_i - \Delta \leq y_i - a_i^T x, & i = 1, 2, \cdots, n; \\ & y_i - a_i^T x \leq t_i + \Delta, & i = 1, 2, \cdots, n; \\ & 0 \leq t_i, & i = 1, 2, \cdots, n. \end{array}$$

The idea of adding a dead zone is to make the large residual relatively more important.

3. If $\rho(x) = |x|$, (7) is the standard least ℓ_1 norm estimation [17]. It can be solved as a linear programming problem [94]. This can be viewed as a special case of the last problem: $\Delta = 0$. This formulation is interesting when the noises are Laplacian: i.e., the errors satisfy an exponential distribution. [61] established an analytical connection between Huber's estimate (with Δ being a tuning parameter) and the least ℓ_1 norm estimate (which was called linear ℓ_1 estimator in [61]). Their result is based on analyzing the solutions to the dual problems, and is inspiring.

5.7.2 Our Choice of Objective Function vs. Huber's Estimate

Our choice of objective function $\rho(\cdot)$ is rooted in (15). We present justification on why to choose such a functional solution as in (16). Because function $-\beta_1 \cdot v(x) - \beta_2$ in (15) is piecewise constant with discontinuity points $-\delta$ and δ, we consider a generic differential equation:

$$\frac{g''}{g} + C = 0, \tag{36}$$

where $C \in \mathbf{R}$ is a constant and $g = \sqrt{f_0}$. The general solution to the above equation, up to a constant, is:

- if $C = 0$, $g = x + c_1$,
- if $C > 0$, $g = \cos(x + c_2)$, and
- if $C < 0$, $g = \exp\{-\sqrt{-C}|x|\}$,

where c_1 and c_2 are constants. Since we want $g(\pm\infty) = 0$, we must assume $-\beta_1 \cdot v(x) - \beta_2 < 0$ outside interval $[-\delta, \delta]$, which leads to the only functional form that vanishes at infinities. Inside interval $[-\delta, \delta]$, we assumed $-\beta_1 \cdot v(x) - \beta_2 > 0$, which leads to the objective function in RSBN. If we choose to assume $-\beta_1 \cdot v(x) - \beta_2 = 0$, then we have $g(x) = x$, which eventually will lead to the Huber's estimate. Our numerical study seems to indicate that our choice leads to relatively more robust performance.

Historically, Huber's estimate is derived differently from our approach, see [60, Section 5.6]. They consider an asymptotic minimax estimate among all cumulative distribution functions (c.d.f.) $F(x) = (1-\varepsilon)G(x) + \varepsilon H(x)$ where constant ε and c.d.f. $G(x)$ are known, and c.d.f. $H(x)$ is unknown but satisfies some general conditions. When $G(x) = \Phi(x)$, which is the c.d.f. of the standard normal, the minimax estimate is the Huber's estimate. Our approach is strongly similar to theirs. However, it differs in the last few steps. We solved the minimax problem in a more general sense.

5.7.3 Other Theoretical Results

Some results that are related to estimators in regression are worth mentioning.

Researchers have explored the robustness of some regression approaches. For example, [26] analyzed the 'leverage' and 'breakdown' in minimum ℓ_1 norm regression. The objective in that paper is different from ours: e.g., they do *not* consider asymptotic performance as we formulated and they do *not* consider a minimax estimate. However, their work is inspiring. A citation search of [26] gives a good sense on what is known about the robustness of some estimators in regression.

In our formulation, we assume the independent noises. Other conditions regarding the regularity of the probability density function of the noises – e.g., the existence of the second derivative of the density, as well as some integrable conditions – are embedded in the derivation of the asymptotic minimaxity. Researchers have studied the condition for an M-estimate to be consistent. The Introduction of a recent article [6] provides a nice overview. Further citation search for the papers cited there gives a full spectrum of the results that are available. In this chapter, we did not intend to address those issues. However, it will be an interesting future search to derive minimax M-estimate under weaker regularity conditions.

5.7.4 Convexity of Fisher Information $I(f)$

In our derivation, we implicitly used the result that Fisher information $I(f)$ is a convex function of f. We give a brief verification of such a convexity. Recall the function $f_t = (1-t)f_0 + tf_1$, $0 \leq t \leq 1$, which was defined in Section 5.3.3. We have

$$I(f_t) = \int \frac{(f_t')^2}{f_t} d\mu(x).$$

One can verify that

$$\frac{\partial^2 I(f_t)}{\partial t^2} \mid_{t=0} = 2 \int \frac{[(f_1' - f_0')f_0 - (f_1 - f_0)f_0']^2}{f_0^3} \geq 0,$$

and the equality is achieved if and only if $f_1 = f_0$, which is *not* true. Hence functional $I(f)$ is strictly convex at every function f_0.

5.7.5 Local Minimaxity

We can only verify that our RSBN estimate is minimax at a neighborhood of function f_0. Reader can refer to Lemma 3.3. Proving that RSBN is a minimax estimate globally (i.e., for all functions satisfying the 'stochastically bounded noise' condition) seems to be a difficult task. This problem has not been solved here.

5.8 Conclusion

We derive an asymptotically minimax estimate in a general regression framework. Extensive numerical simulation demonstrates its advantage over ordinary least square estimate, as well as another robust estimate: Huber's M-estimate.

An interesting insight of our result is to observe that the derived objective function should be in the form of the ℓ_1 norm outside a neighborhood of the origin. This coincides with many recent applications of ℓ_1 norm in problems such as variable selection. Even though this chapter does not exactly create any link, the connection between the ℓ_1 norm and the asymptotic minimaxity of RSBN is certainly something that should be explored in the future.

A condensed version of this chapter can be seen in [72].

CHAPTER VI

ACHIEVING OPTIMAL REPRESENTATION WITH STEPWISE ALGORITHMS IN REGRESSION

This chapter presents new results on using stepwise algorithm to achieve the best representations of signals that coincide with model selection results. This is motivated first by the analysis on the performance of a newly developed algorithm, least angle regressions (LARS). A counter example is established to show that LARS cannot recover the optimal selection in certain cases. Conditions under which LARS (Lasso) or stepwise algorithms can recover exactly the optimal models are investigated. We study the homotopy between LARS and Lasso and reveal that LARS yields the Lasso solution path. This is a known result in the literature [25]. These problems, which are raised in Lasso and LARS, are outlined with **(P1)**. Meanwhile, Classical model selection criteria are reviewed, which are summarized with **(P0)**. Problem **(P0)** is combinatorial in nature and proven to be NP-hard. We try to investigate the relationship between **(P0)** and **(P1)** and hence find the connection between stage-wise algorithms and statistical variable selection critera. Several conditions are given. We present the necessary and sufficient condition for a vector to be the optimal solution of **(P1)**. For **(P0)**, sufficient conditions are derived. We also study the conditions under which the two optimization problems have common solutions. Hence, in these situations, a greedy algorithm can be used to solve the seemingly unsolvable problem. We provide the results from three different angles: (1) a direct analysis on sufficiency and necessity, (2) results on covariates that are mostly correlated with the response, (3) results motivated by recent works in sparse signal representation. The applications, possible future research, and related works in statistics are discussed.

This chapter is organized as follows. Section 6.1 introduces the two optimization problems we are considering, **(P0)** and **(P1)**, together with their connection with modern subset selection criteria, Lasso, and LARS. Section 6.2 reviews the known model selection criteria in statistics, as well as the solution paths property of Lasso and its solutions based on LARS. This material provides a starting point of the consequent work. Section 6.3 presents two case studies. In the first case, it is shown that a greedy algorithm (i.e., a version of LARS) can go totally wrong in an extreme situation. In the second case, it is shown that the two optimization problem we are considering give the same result in subset selection. These two opposing cases motivate us to analyze the conditions under which the two approaches choose the identical subset. Section 6.4 contains the main results. Our main results are organized in three groups. In Section 6.4.1, necessary and sufficient conditions are provided. For **(P0)**, such a condition is hard to verify in practice. In Section 6.4.2, a sufficient condition is derived. This condition started from a simple fact: the most correlated covariates (with the response) form the concurrent optimal subset. This condition is easy to verify numerically. However, it is relatively restrictive. We use it as a preparation for more flexible sufficient conditions. In Section 6.4.3, a very general sufficient condition is derived. To our knowledge, this is the best known subset equivalence condition between **(P0)** and **(P1)**. Section 6.5 discusses related works and potential future research topics. A brief conclusion is provided in Section 6.6. To keep the flow of the paper, not-directly-required proofs are postponed into the appendix B.

6.1 Introduction

We consider two types of optimization problems in this chapter.

- The first is an optimization problem that is based on a counting measure,

$$\textbf{(P0)} \qquad \min_{x} \quad \|y - \Phi x\|_2^2 + \lambda_0 \cdot \|x\|_0,$$

  where $\Phi \in \mathbb{R}^{n \times m}, x \in \mathbb{R}^m, y \in \mathbb{R}^n$, the notation $\| \cdot \|_2^2$ denotes the sum of squares of the entries of a vector, nonnegative constant λ_0 is an algorithmic parameter, and the quantity $\|x\|_0$ is the number of nonzero entries in vector x.

- Solving **(P0)** generally requires exhaustive searching through of all the possible subsets. When m, the column size of Φ, increases, the methods based on exhaustive search become rapidly impractical. An approach is to relax the above problem by replacing

$\|x\|_0$ with $\|x\|_1$, which leads to the following problem: an optimization problem that depends on a sum of absolute values,

$$\textbf{(P1)} \qquad \min_{x} \quad \|y - \Phi x\|_2^2 + \lambda_1 \cdot \|x\|_1,$$

where $\|x\|_1 = \sum_{i=1}^{m} |x_i|$ for vector $x = (x_1, x_2, \ldots, x_m)^T$, and the nonnegative constant λ_1 is another algorithmic parameter, whose role will be discussed later.

Note that $\|x\|_0$ (respectively, $\|x\|_1$) is a quasi-norm (respectively, norm) in $\mathbb{R}^m$. In the literature of *sparse signal presentation*, they are called the ℓ_0-norm and the ℓ_1-norm, respectively. The numbers "0" and "1" in the notations **(P0)** and **(P1)** follow such a convention [20, 19, 11].

In subset selection under linear regression, many well known criteria – including C_p statistic [65], Akaike information criterion (AIC) [1], Bayesian information criterion (BIC) [84], minimum description length (MDL), risk inflation criterion (RIC) [30], and so on – are special cases of **(P0)**, by assigning different values to λ_0. Details regarding the foregoing statement will be provided later. It is shown in this paper that the problem **(P0)** in general is NP-hard (Theorem 2.1).

At the same time, **(P1)** is the mathematical problem that is called upon in Lasso [90]. Recent advances (whose details and references are provided in Section 6.2.2) demonstrate that some stepwise algorithms (e.g., least angle regressions (LARS) presented in [25]) reveal the solution paths of problem **(P1)**, while parameter λ_1 takes a range of values. More importantly, most of these algorithms only take polynomial number of operations – i.e., they are polynomial-time algorithms and **(P1)** minimizes a global objective function. In fact, the complexity of finding a solution path for **(P1)** is the same as implementing an ordinary least square fit [25].

However, as pointed out by the authors of LARS, having the same solutions as Lasso does not guarantee that LARS selects the optimal subset of variables. I.e., solutions to **(P1)** may not match the principles of AIC, BIC, C_p, and RIC **(P0)**. An example by Weisberg in the discussion of LARS ([95]) is unfavorable to the optimality of LARS. This chapter will also give an example in which stepwise algorithms go totally wrong until the last step. I.e.

the algorithm chooses all the variates outside the optimal subset before it selects any inside the optimal subset.

Therefore, the issue of under what conditions a stepwise greedy approach can generate a solution that optimizes a global objective function interests us. For clarity, we restricted the underlying model to be a linear regression model. Variable selection instead of model selection is the focus: we are *not* giving an optimality criterion for model selection; instead, conditions under which **(P0)** and **(P1)** give the same result are investigated.

In summary, the main objective of this paper is to find when **(P0)** and **(P1)** lead to a common solution in the subset selection under a regression model. A subset that corresponds to the nonzero subset of the minimizer of **(P0)** (respectively, **(P1)**) is called a *type-I* (respectively, *type-II*) *optimal subset with respective to* λ_0 (respectively, λ_1). A subset that is both type-I and type-II optimal is called a *concurrent optimal subset.* It will be shown that there is a necessary and sufficient condition for the type-II optimal subset (Theorem 4.2), and this condition can be verified in polynomial time. However, in general, there is no polynomial-time necessary and sufficient condition for the type-I optimal subset. We then search for easy-to-verify (i.e., polynomial-time) sufficient conditions for type-I optimal subsets. Two types of results are derived. The first is based on the assumption that the most correlated covariates form the optimal subset. The second result is motivated by a new advance in sparse signal representation, and is rather general.

Our analysis deals with a fundamental issue that has recurred due to the introduction of LARS. In practice, stepwise greedy algorithms are normally preferred by empiricists due to their simplicity in implementation, while global optimality criteria are favored by theorists due to their amenability in analysis. Theories explaining the link between these two is of both practical and theoretical interests. Our work will raise awareness, and more importantly, encourage more research on this topic.

6.2 Review of Two Optimization Problems

We consider *subset selection* in regression. Recall in a regression setting, $\Phi \in \mathbb{R}^{n\times m}(n > m)$ denotes a model matrix. Vectors $x \in \mathbb{R}^m$ and $y \in \mathbb{R}^n$ are coefficient and response vectors.

The columns of matrix Φ are *covariates*. A regression model is $y = \Phi x + \varepsilon$, where ε is a random vector. Let $\mathbf{I} = \{1, 2, \ldots, m\}$ denote all the indices of the coefficients. A subset of coefficients (or, covariates) is denoted by $\Omega \subseteq \mathbf{I}$. Let $|\Omega|$ denote the cardinality of the set Ω. Let x_Ω denote the coefficient vector that only takes nonzero values when the coefficient indices are in the subset Ω. A subset selection problem has two competing objectives in choosing a subset Ω: firstly, the residuals, which are in the vector $y - \Phi x_\Omega$, are close to zeros; secondly, the size of the set Ω is small. Note that we differ from many statisticians, who emphasize the predictability of the selected models. We provide some discussions in Section 6.5.

6.2.1 Subset Selection Criteria and (P0)

There has been rich literature on the criteria regarding subset selection. Book [68] and paper [31] give an excellent overview. An interesting fact is that a majority of these criteria can be unified under **(P0)**, where $\|y - \Phi x\|_2^2$ is the residual sum of squares (denoted by RSS(x)) under the coefficient vector x, and constant λ_0 depends on the criteria. The following summarizes some well-known results:

- Akaike [1] defines his criterion by maximizing the expected log-likelihood $E_{X,\hat{\theta}}(\log f(X|\hat{\theta}))$, where $\hat{\theta}$ is the estimate of parameter θ, $f(X|\theta)$ is the density function. This is equivalent to maximizing the expected Kullback-Leibler's mean information for discrimination between $f(X|\hat{\theta})$ and $f(X|\theta)$, i.e., $E_{X,\hat{\theta}}(\log \frac{f(X|\hat{\theta})}{f(X|\theta)})$, for a known true θ. Under a Gaussian assumption in the linear regression, the above leads to the Akaike information criterion (AIC) that minimizes

$$\text{AIC} = \frac{\text{RSS}(x)}{\sigma^2} + 2 \cdot \|x\|_0,$$

  where σ^2 is the noise variance, and other notations have been defined at the beginning of this section. It is a special case of **(P0)** by assigning $\lambda_0 = 2\sigma^2$.

- Mallows' C_p [65, 34], which is derived from the unbiased risk estimation, minimizes

$$C_p = \frac{1}{\hat{\sigma}^2} RSS(x) + 2 \cdot \|x\|_0 - n,$$

where $\hat{\sigma}$ is an estimate of the parameter σ. When $\hat{\sigma}^2 = \sigma^2$ is assumed, the C_p is equivalent with the AIC. Again C_p is a special case of **(P0)**.

- Motivated by the asymptotic behavior of Bayes estimators, Bayesian information criterion (BIC) [84] chooses to select the model that maximizes

  $$\log f(X|\hat{\theta}) - \frac{1}{2} \cdot \log n \cdot \|x\|_0.$$

  Again, under the squared error loss and the Gaussian model assumption with known variance σ^2, BIC is to minimize

  $$\text{BIC} = \frac{RSS(x)}{\sigma^2} + \log n \cdot \|x\|_0.$$

  The above is a special case of **(P0)** by assigning $\lambda_0 = \sigma^2 \log n$.

- According to [41, Section 7.8], the equivalence between BIC and the minimum description length (MDL) is well known. Hence MDL is a special case of **(P0)**.

- Risk inflation criterion (RIC) is suggested in [30] from a minimax estimation vantage point. RIC recommends the model that minimizes

  $$\text{RIC} = \frac{RSS(x)}{\sigma^2} + 2\log p \cdot \|x\|_0,$$

  where p is the number of available predictors. This is derived from selecting the model with minimum risk inflation. Due to the different emphasis of the present paper, we do not include further details of RIC. However, readers can see that RIC is another special case of **(P0)**, by taking $\lambda_0 = 2\sigma^2 \log p$.

In this paper, the "subset selection criteria" that appears everywhere encompasses all the aforementioned criteria, all adopting the formulation **(P0)**.

Solving **(P0)** generally requires exhaustive search of all the possible subsets. When $\|x\|_0$ (i.e., the number of covariates) increases, the methods based on exhaustive search become rapidly impractical. In fact, solving **(P0)** in general is an NP-hard problem. The following theorem can be considered as an extension of a result that was originally presented in [71].

Theorem 2.1 *Solving the problem* **(P0)** *with a fixed* λ_0 *is an NP-hard problem.*

Proof. Let

$$f(m) = \min_{x:\ \|x\|_0 \le m} \quad \|y - \Phi x\|_2^2,$$

where all the symbols are defined in **(P0)**. It is evident that point array $(m, f(m)), m = 1, 2, \ldots$, forms a non-increasing curve in the positive quadrant.

We first establish the existence of an integer m_0, such that value $f(m_0)+\lambda_0 m_0$ minimizes the objective in **(P0)**. Note that there are finite number of m's such that $\lambda_0 m \le f(1)+\lambda_0 \cdot 1$. This inequality gives an upper bound of m's that satisfy $f(m)+\lambda_0 m \le f(1)+\lambda_0 \cdot 1$. Among these finite number of m's, there is at least one m_0 that minimizes the value of function $f(m) + \lambda_0 m$.

Define $\varepsilon = f(m_0)$. In general, we can assume $\varepsilon > 0$, because if $\varepsilon = 0$, response y can be superposed by a small (more specifically, no more than m_0) number of columns of matrix Φ, which is a special case.

Using the idea of Lagrange multiplier, we can see that solving **(P0)** with λ_0 is equivalent to solving the sparse approximate solution (SAS) problem in [71, Section 2] with ε, which is proven in [71] to be NP-hard. Hence, in general, solving **(P0)** is NP-hard. □

6.2.2 Greedy Algorithms and (P1)

Due to the hardness of solving **(P0)**, a *relaxation* idea has been proposed. The relaxation replaces the ℓ_0 norm with the ℓ_1 norm in the objective, which leads to **(P1)**. The idea of relaxation started in *sparse signal representation* [12]. Theoretical properties are derived later in [20, 19]. A partial list of new representative results include [91], [92], [37], and [11]. Being compared with this paper, the problem of sparse signal representation has a different emphasis. In sparse signal representations, researchers consider a redundant *dictionary* [63, 32] and the conditions under which the sparsest representation can be solved via a linear programming. Their formulations of **(P0)** and **(P1)** are slightly different from ours. However, a group of results in this paper are certainly motivated by some recent results in sparse representation. More connections will be discussed when we present our findings in Section 6.4.3.

At the same time, **(P1)** has been proposed in statistics as a way of subset selection. The method is coined as Lasso [90]. An interesting recent development – the least angle regressions (LARS) [25] – demonstrates that certain greedy algorithms can reveal the solutions to **(P1)** with varying values of λ_1, based on the idea of *homotopy* [77]. Here, we review this result by a simple illustration. We prove that LARS is derived by satisfying a necessary condition for a vector being an optimal solution in Lasso (i.e. **(P1)**), which represents the idea of homotopy used in the LARS paper [25]. Being compared with those existing homotopy explanations in [77], the following analysis is more straightforward, taking advantage of a Lagrange multiplier and a perturbation analysis.

Recall Lasso is equivalent to find the following minimizer:

$$\begin{aligned} x(c) = \quad & \text{argmin} \quad \|y - \Phi x\|_2^2, \\ & \text{s.t.} \quad \|x\|_1 \leq c, \end{aligned} \tag{37}$$

where c is a constant; y, Φ, x, and $\|x\|_1$ have been defined before. The sum of squares of residuals is $\|y - \Phi x\|_2^2$. To make a link later, the following graphical illustration of (37) is introduced. In Fig. 55, the horizontal axis is the value of $\|x\|_1$, the vertical axis is the

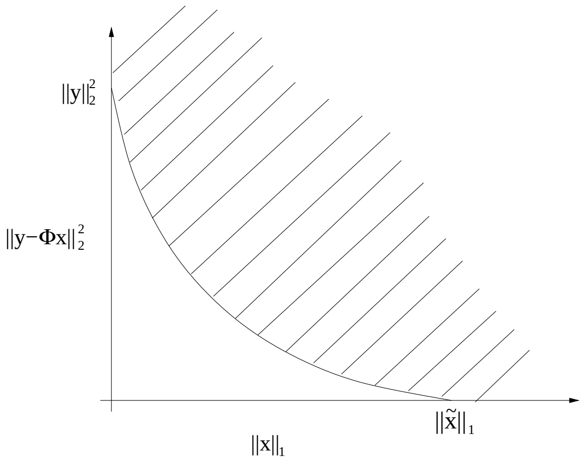


Figure 55: Graphical Illustration of Lasso Problem.

value of $\|y - \Phi x\|_2^2$. The point set $(\|x\|_1, \|y - \Phi x\|_2^2)$, for all c, forms a feasible set. This set is the shaded region in Fig. 55. The lower bound of the feasible set is called a *frontier*. Apparently, each point on the frontier corresponds to a solution to (37), with a particular constant c. Given that function $\|y - \Phi x\|_2^2$ is strictly convex for x, one can verify that the frontier in this case is strictly convex. This intuitively correct phenomenon is hard to be proved. It is listed as a lemma below. In Fig. 55, $\tilde{x}$ denotes a solution of $0 = y - \Phi x$.

Lemma 2.2 *If there exists a vector $\tilde{x}$, such that $0 = y - \Phi\tilde{x}$, then the frontier mentioned above is strictly convex, i.e., there are* **no** *vectors x_1 and x_2 and constant $\lambda (0 < \lambda < 1)$, such that points $p_1 = (\|x_1\|_1, \|y - \Phi x_1\|_2^2)$, $p_2 = (\|x_2\|_1, \|y - \Phi x_2\|_2^2)$, and $\lambda p_1 + (1 - \lambda)p_2$ are simultaneously on the frontier.*

The proof can be found in Appendix B.1.

Now we apply the idea of Lagrange multipliers. For every c, there exists a value λ, such that

$$x(c) = x(\lambda) = \text{ argmin } \quad \|y - \Phi x\|_2^2 + \lambda \|x\|_1. \tag{38}$$

This indicates that Lasso solves **(P1)**. Being compared to (37), the optimization problem (38) is unconstrained. Hence, we can consider the First Order Condition for the objective in (38). Let $f(x;\lambda)$ denote the objective:

$$f(x;\lambda) = \|y - \Phi x\|_2^2 + \lambda \|x\|_1.$$

We have the first derivative

$$\begin{aligned} \frac{\mathrm{d}f(x;\lambda)}{\mathrm{d}x} &= 2\Phi^T \Phi x - 2\Phi^T y + \lambda \cdot \text{sign}(x) \\ &= -2\Phi^T (y - \Phi x) + \lambda \cdot \text{sign}(x), \end{aligned}$$

where sign(x) is a vector whose entries are the signs of the entries of vector x. The above is written assuming x_i's are not equal to zero. When x_i is zero, $f(x;\lambda)$ is not differentiable.

From the above, we have

$$\frac{\mathrm{d}f(x;\lambda)}{\mathrm{d}x_i} = -2\left[\Phi^T (y - \Phi x)\right]_i + \lambda \cdot \text{sign}(x_i)$$

for $x_i \neq 0$. Hence

$$\left[\Phi^T (y - \Phi x)\right]_i = \frac{\lambda}{2} \cdot \text{sign}(x_i). \tag{39}$$

If $x_i = 0$, and x minimizes $f(x;\lambda)$, we must have

$$\left|-2\left[\Phi^T (y - \Phi x)\right]_i\right| \leq \lambda;$$

otherwise, a small perturbation of x_i will decrease the value of $f(x;\lambda)$. Hence

$$\left|\left[\Phi^T(y-\Phi x)\right]_i\right| \le \lambda/2; \tag{40}$$

From the above, if x minimizes $f(x;\lambda)$, both (39) (when $x_i \neq 0$) and (40) (when $x_i = 0$) must be satisfied. On the other hand, if x satisfies (39) and (40), then x is at least a local minimizer of $f(x;\lambda)$. Note that even the frontier is strictly convex, the minimizer of the function f is not necessarily unique. For example, there exist v_1 and v_2 such that $\Phi(v_1 - v_2) = 0$ and $\text{sign}(v_1) = \text{sign}(v_2)$. One can verifies that if $f(v_1;\lambda) = f(v_2;\lambda)$, then $f(v_1;\lambda) = f(\kappa v_1 + (1-\kappa)v_2;\lambda)$, for $0 \le \kappa \le 1$.

According to the steps of LARS and the above analysis, we proved the following.

Theorem 2.3 *At each iteration of LARS, a solution vector satisfies a necessary condition for this vector to be a solution to Lasso.*

More recent analysis demonstrates further that greedy algorithms can literally render the entire solution path in a large class of problems, referring to [39] and the references therein. A recent conference presentation [62] gives the most succinct solution in generating solution paths, utilizing a homotopy continuation method [78] and an analysis of *subdifferential.* [83] is a standard reference for the background of this material.

6.3 Motivations: Case Studies

6.3.1 An Extremal Example for the Least Angle Regressions

Least Angle Regression [25] is a forward variable selection method. An extensive manual regarding forward selection can be found in [5]. As been indicated previously, LARS can give the solution path of **(P1)**. However, this homotopy does not guarantee that LARS always reveal the optimal solutions of **(P0)**. In this subsection, we present one particular case, in which LARS choose wrongly in the first iteration and end up correcting it inefficiently. As a result, LARS do not include the correct covariates until the last step. Initially, such an example motivated us to consider the conditions that will be presented later.

Details of LARS algorithm can be found in [25], Section 2. In simplicity, LARS start with zero coefficients, select the most correlated covariates with the signal s, then move along the direction that is equiangular among the selected covariates until some other covariates have as much correlation with the current residual, add these new covariates under consideration and move along the new equiangular direction. When the covariates and the response are standardized to have mean 0 and unit norm, correlation between vectors is proportional to the inner product. In this section, for clarity, we first give an example with nonstandardized vectors, and choose the covariates according to the inner products. The corresponding example with standardized covariates and signal is presented later in Section 6.3.1.1. Section 6.3.1.2 shows how to use the result in this section to come up with a dramatic example in presentation.

The first example is generated as follows. Let $\phi_i \in \mathbb{R}^n, i = 1, 2, ..., m$, denote the ith column of the model matrix Φ. Hence, $\Phi = [\phi_1, \phi_2, ..., \phi_m]$. Let $\delta_i \in \mathbb{R}^n, i = 1, 2, ..., m$, denote the dirac vector taking 1 at the ith position and zero elsewhere. For $i = m - A + 1, m - A + 2, ..., m$, let $\phi_i = \delta_i$, where A is a positive integer. Consider a special signal $s = \frac{1}{\sqrt{A}} \sum_{i=m-A+1}^{m} \phi_i$. Obviously, in this case, the optimal subset is $\{m - A + 1, ..., m\}$. For the first $m - A$ columns of Φ, make $\phi_j = a_j \cdot s + b_j \cdot \delta_j$, where $1 \leq j \leq m - A$ and $a_j^2 + b_j^2 = 1$. Note ϕ_i's and s are all unit-norm vectors. From now on, for simplicity, we always assume $1 \leq j \leq m - A$ and $m - A + 1 \leq i \leq m$. It is easy to verify that

$$\langle s, \phi_j \rangle = a_j \qquad \text{and} \qquad \langle s, \phi_i \rangle = 1/\sqrt{A}.$$

In this example, we choose $1 > a_1 > a_2 > \cdots > a_{m-A} > 1/\sqrt{A} > 0$.

Now consider the procedure of LARS. In the first step, since ϕ_1 has the largest inner product with s, evidently column ϕ_1 will be chosen. The next residual will be $r_1 = s - c_1\phi_1$, where c_1 is the coefficient to be determined. The following result about the consequent step in LARS will be proved in Appendix B.2.

Lemma 3.1 *In the consequent step of LARS, covariate ϕ_2 is chosen, with $c_1 = \frac{a_1 - a_2}{1 - a_1 a_2}$.*

Hence, the residual of the first step becomes

$$\begin{aligned} r_1 &= s - c_1\phi_1 \\ &= s - \frac{a_1 - a_2}{1 - a_1a_2}(a_1 s + b_1\delta_1) \\ &= \frac{b_1^2}{1 - a_1a_2}s - \frac{(a_1 - a_2)b_1}{1 - a_1a_2}\delta_1 \\ &= \frac{b_1^2}{1 - a_1a_2}[s - \frac{a_1 - a_2}{b_1}\delta_1]. \end{aligned}$$

Note that in LARS, only the direction of a residual vector determines the selection of the next covariates. The amplitude of a residual vector does not change the variable selection. Hence, we introduce a surrogate residual with a simpler form:

$$\widetilde{r}_1 = s - \frac{a_1 - a_2}{b_1}\delta_1.$$

Residuals $\widetilde{r}_1$ and r_1 have the same direction. This is an important step to simplify our analysis. In the proof of the next theorem, the surrogate residuals with simpler forms are repeatedly called upon.

As a sanity check, the following calculations are performed:

1. For i, $\langle \phi_i, \widetilde{r}_1 \rangle = 1/\sqrt{A}$.

2. For j,

$$\begin{aligned} \langle \phi_j, \widetilde{r}_1 \rangle &= \langle a_j s + b_j\delta_j, s - \frac{a_1 - a_2}{b_1}\delta_1 \rangle \\ &= a_j - \frac{b_j(a_1 - a_2)}{b_1}\langle \delta_j, \delta_1 \rangle. \end{aligned}$$

As special cases: $\langle \phi_1, \widetilde{r}_1 \rangle = a_2$, $\langle \phi_2, \widetilde{r}_1 \rangle = a_2$, and for $j \geq 3$, $\langle \phi_j, \widetilde{r}_1 \rangle = a_j$.

The above analysis demonstrates some basic techniques that will be used in the consequent LARS steps. Now we use induction to show the following.

Theorem 3.2 (Case Study of LARS) *In the example described in the beginning of this section, LARS choose covariates $\phi_1, \phi_2, ..., \phi_{m-A}$ one by one sequentially in the first $m - A$ steps.*

It takes some energy to verify the above theorem. We postpone it to Appendix B.3. This example shows that LARS can choose all the covariates outside an intuitively optimal subset before it reaches any covariate inside the optimal subset.

6.3.1.1 Standardized Covariates

Readers may notice that LARS should proceed along the direction that depends on the correlations between ϕ_i's and the residual. Meanwhile, in our case study, the proceeding direction is determined due to the inner product. The inner product is not proportional to the correlation since the response s and the covariate vectors ϕ_i's are not standardized to have mean 0. However, this discrepancy can be easily remedied as follows. The key observation is that LARS only depend on geometric information. More specifically, the result depends only on $\langle \phi_i, s \rangle$, $i = 1, 2, ..., m$, and $\langle \phi_i, \phi_j \rangle$, $1 \leq i, j \leq m$. For example, an orthogonal transform of both s and ϕ_i's will retain the results in LARS. We state this without a proof.

Lemma 3.3 *After a simultaneously orthogonal transform on both response and covariates, the results of LARS from the transformed data is the same orthogonal transform of the LARS results from the original data.*

Hence, if we can find another set of standardized vectors, which retain the inner products and are the orthogonal transforms of ϕ_i's and s in the previous example, the same results can be predicted for LARS.

The standardization can be incorporated according to the following. The main idea is that an n-dimensional linear space can be treated as a subspace of $\mathbb{R}^{n+1}$, which is orthogonal to vector $(1, 1, ..., 1)$. Let $\{b_0, b_1, ..., b_n\}$ denote an orthonormal basis of $\mathbb{R}^{n+1}$, with $b_0 = \frac{1}{\sqrt{n+1}}(1, 1, ..., 1)^T$. Denote the unit-norm vectors $s = (s_1, s_2, ..., s_n)^T$ and $\phi_i = (\phi_{i1}, \phi_{i2}, ..., \phi_{in})^T$, $i = 1, 2, ..., m$. Define $s' = \sum_{j=1}^n s_j b_j$, $\phi_i' = \sum_{j=1}^n \phi_{ij} b_j$, $i = 1, 2, ..., m$. One can easily verify that $\langle s', \phi_i' \rangle = \langle s, \phi_i \rangle$ for $1 \leq i \leq m$, and $\langle \phi_i', \phi_j' \rangle = \langle \phi_i, \phi_j \rangle$ for $1 \leq i, j \leq m$. Hence, applying LARS to s' and ϕ_i''s will produce the same result as in the first case study. It is not hard to verify that s' and ϕ_i''s are standardized. Hence, the conclusions in our case study can be extended to the case with standardized response and covariates.

Theorem 3.4 (An Example with Standardized Covariates) *There exists an orthogonal transform that can be applied to the previous example to create a case in which all the covariates and the response are standardized, and LARS select all the covariates outside the optimal subset before it chooses any covariate inside the optimal subset.*

6.3.1.2 A Dramatic Presentation

The foregoing example is developed in a fairly general form, with controlling parameters A and m. To illustrate how dramatic this example can be, let us consider the case where $A = 10$ and $m = 1,000,000$. Based on the previous description, the LARS will select the first $999,990$ covariates before it selects any of the last ten covariates. At the same time, the optimal subset is formed by the last ten covariates.

6.3.2 Variable Selection with Orthogonal Model Matrix

In order to provide some insights, a simple case in which Φ is orthogonal is considered. Although this example has been studied in the original LARS paper [25], the purpose of restating it here is to illustrate that there is a case in which LARS find the type-I optimal subset.

Theorem 3.5 (Orthogonal Design) *Let $\widetilde{x}_0$ and $\widetilde{x}_1$ denote the solutions to* **(P0)** *and* **(P1)***, respectively. When Φ is orthogonal, we have*

$$\widetilde{x}_{0,i} = \begin{cases} 0, & \textit{if} \quad |z_i| \le \sqrt{\lambda_0}, \\ z_i, & \textit{if} \quad |z_i| > \sqrt{\lambda_0}, \end{cases}$$

and

$$\widetilde{x}_{1,i} = \begin{cases} 0, & \textit{if} \quad |z_i| \le \lambda_1/2, \\ sign(z_i)(|z_i| - \frac{\lambda_1}{2}), & \textit{if} \quad |z_i| > \lambda_1/2. \end{cases}$$

Here, $\widetilde{x}_{0,i}$ and $\widetilde{x}_{1,i}$ denote the ith entry of $\widetilde{x}_0$ and $\widetilde{x}_1$, respectively, and z_i is the ith entry of $z = \Phi^T y$.

For readers who are familiar with soft-thresholding and hard-thresholding [23], the above is not a surprise. A proof follows.

Proof. Both **(P0)** and **(P1)** can be decomposed into the univariate problems

$$\min_{x_i} \quad (z_i - x_i)^2 + \lambda_0 \cdot 1(x_i \neq 0),$$

and

$$\min_{x_i} \quad (z_i - x_i)^2 + \lambda_1 \cdot |x_i|.$$

From here, it is not hard to derive the formulae in the theorem. □

From the above, verifying the following becomes an easy task. Let supp(x) denote the set of indices of the nonzero entries in vector x.

Corollary 3.6 *When $\sqrt{\lambda_0} = \lambda_1/2$, one has supp($\widetilde{x}_0$) = supp($\widetilde{x}_1$), i.e., there is a concurrent optimal subset. Moreover,*

$$\widetilde{x}_{0,i} - \widetilde{x}_{1,i} = \begin{cases} 0, & \textit{if} \quad i \notin \textit{supp}(\widetilde{x}_0), \\ \frac{\lambda_1}{2} \cdot \textit{sign}(z_i), & \textit{if} \quad i \in \textit{supp}(\widetilde{x}_0). \end{cases}$$

The proof is obvious and is omitted.

Now there are two opposing examples. On one hand, if Φ is orthogonal, LARS and Lasso recover the optimal subset in **(P0)**. On the other hand, we found an example in which a version of LARS would choose all the covariates outside the optimal subset before choosing anything inside. These inconsistencies encourage us to analyze the solutions of **(P0)** and **(P1)**, and the conditions for a subset to be the concurrent optimal subset. We present the details and the results in the next section.

6.4 Main Results: Conditions of Equivalence

We present our findings in three subsections. In Section 6.4.1, we give a sufficient and necessary condition for a subset to be the concurrent optimal subset. Recall that **(P0)** in general is NP-hard. Checking the aforementioned condition can not be done via a polynomial-time algorithm. In Section 6.4.2, we ask when the k most correlated covariates form the concurrent optimal subset. A sufficient condition is derived. This result is easy to check but too restrictive. However, it inspires us to consider more general sufficient conditions. A more general sufficient condition for **(P0)** is derived in the next section – Section 6.4.3 – which is also motivated by a recent approach appeared in applied mathematics [36]. We modified their approach to solve a different mathematical problem.

6.4.1 Sufficient and Necessary Conditions

Before moving into the specific discussion, we introduce a sufficient and necessary condition for a concurrent optimal subset. Let I_1 denote a subset of indices. Let Φ_1 and x_1 denote columns of Φ and entries of x with indices from I_1. Let $\Phi = [\Phi_1 \ \Phi_2]$. Here, a permutation that does not change the problem is implied.

Theorem 4.1 (Sufficient and Necessary for (P0)) *I_1 is the optimal subset of* **(P0)** *if and only if value*

$$y^T y - y^T \Phi_1 (\Phi_1^T \Phi_1)^{-1} \Phi_1^T y + \lambda_0 \cdot \|x_1\|_0 \tag{41}$$

is the minimum of the objective in **(P0)***.*

Theorem 4.2 (Sufficient and Necessary (P1)) *I_1 is the optimal subset of* **(P1)** *if and only if there exists a vector ω, such that*

$$\Phi^T y = \begin{pmatrix} \Phi_1^T \Phi_1 \\ \Phi_2^T \Phi_1 \end{pmatrix} x_1 + \begin{pmatrix} \frac{\lambda_1}{2} \cdot sign(x_1) \\ \omega \end{pmatrix} \tag{42}$$

holds and $\|\omega\|_\infty \leq \lambda_1/2$.

Theorem 4.3 (Sufficient and Necessary (Concurrent)) *I_1 is the concurrent optimal subset of* **(P0)** *and* **(P1)** *if and only if (41) and (42) are true. Moreover, recall $\widetilde{x}_0$ and $\widetilde{x}_1$ are the solutions of* **(P0)** *and* **(P1)***, respectively. We have*

$$(\widetilde{x}_0 - \widetilde{x}_1)_{I_1} = (\Phi_1^T \Phi_1)^{-1} \cdot \frac{\lambda_1}{2} \cdot sign((\widetilde{x}_1)_{I_1}). \tag{43}$$

Proof. For the above theorems, Theorem 4.1 is from a direct derivation; and Theorem 4.2 is based on the argument of *subdifferential* [92, 62].

For Theorem 4.3, consider

$$(\widetilde{x}_0)_{I_1} = (\Phi_1^T \Phi_1)^{-1} \Phi_1^T y,$$

and

$$\Phi_1^T y = (\Phi_1^T \Phi_1)(\widetilde{x}_1)_{I_1} + \frac{\lambda_1}{2} \cdot \mathrm{sign}((\widetilde{x}_1)_{I_1}).$$

By combining the above two, (43) follows. □

The above theorem gives a necessary and sufficient condition for a concurrent optimal subset. The following provides some further comments.

Remark 4.4 *Equation (43) provides a methods of computing $\widetilde{x}_1$, given that $\widetilde{x}_0$ is available and represents the optimal solution. Evidently,*

$$(\widetilde{x}_1)_{I_1} = (\widetilde{x}_0)_{I_1} - \frac{\lambda_1}{2}(\Phi_1^T \Phi_1)^{-1} \cdot \mathit{sign}((\widetilde{x}_1)_{I_1}).$$

Remark 4.5 *Note*

$$\begin{aligned} \Phi(\widetilde{x}_0 - \widetilde{x}_1) &= \Phi_1(\widetilde{x}_0 - \widetilde{x}_1)_{I_1} \\ &= \frac{\lambda_1}{2} \cdot \Phi_1(\Phi_1^T \Phi_1)^{-1} \cdot \mathit{sign}((\widetilde{x}_1)_{I_1}), \end{aligned}$$

which is an equiangular vector among the columns of Φ_1. Hence, when optimality is achieved in both (42) and (43), the difference between the two predicted vectors is an equiangular vector.

Readers can compare the above results with those in [62], who independently achieved the same results.

Verification of the sufficient and necessary conditions in (41) is difficult, requiring solving a combinatorial search problem. Because in general, solving **(P0)** is NP-hard (Theorem 2.1), it will be easy to verify that there should be no sufficient and necessary condition that can be verified by a polynomial time algorithm.

6.4.2 A Sufficient Condition for Covariates that are Mostly Correlated with the Response

Because it is generally impossible to have a necessary and sufficient condition that can be verified in polynomial time, we will focus on finding some easy-to-verify sufficient conditions.

We first introduce a set of sufficient conditions, which only depend on the correlations between the response y and the covariates ϕ_i, as well as the maximum correlation between the covariates. For simplicity, we now assume that response y and covariates ϕ_i's are all standardized. It is not hard to see $|\langle y, \phi_i \rangle| \leq 1$, $i = 1, 2, ..., m$, and $|\langle \phi_i, \phi_j \rangle| \leq 1, 1 \leq i, j \leq m$. Denote $z = \Phi^T y = (z_1, z_2, ..., z_m)^T$. Without loss of generality, we assume $|z_1| > |z_2| > \cdots > |z_m|$. We want to find sufficient conditions such that subset $A_1 = \{\phi_1, \phi_2, ..., \phi_k\}$ is the solution to both **(P0)** and **(P1)**. In other words, the k most correlated covariates with the response form the optimal subset. Clearly, an optimal subset does not need to be the most correlated covariates with the response. Due to this additional condition, this set of conditions are *restrictive.* The restrictiveness is illustrated in an example in Section 6.4.2.1.

Denote

$$\mu = \max_{\substack{1 \leq i,j \leq m \\ i \neq j}} |\langle \phi_i, \phi_j \rangle|.$$

The following is a well-known result from linear algebra.

Lemma 4.6 *Let $\lambda(\Phi_1^T \Phi_1)$ denote an eigenvalue of matrix $\Phi_1^T \Phi_1$, where $\Phi_1 = [\phi_1, \phi_2, \cdots, \phi_k]$. We have*

$$1 - (k-1)\mu \leq \lambda(\Phi_1^T \Phi_1) \leq 1 + (k-1)\mu, \tag{44}$$

and

$$\frac{1}{1 + (k-1)\mu} \leq \lambda\left((\Phi_1^T \Phi_1)^{-1}\right) \leq \frac{1}{1 - (k-1)\mu} \tag{45}$$

The above lemma will be used in proving the following two theorems.

Theorem 4.7 *For a given λ_0, and correlations $z_1, z_2, ..., z_k$, if the following three conditions are satisfied:*

$$[1 - (k-1)\mu] z_k^2 \geq 2(k-1)^2 \mu + z_{k+1}^2 [1 + (k-1)\mu], \tag{46}$$

$$z_{k+1}^2 \leq \lambda_0 (1 - \Delta) - \frac{(2k-1)\mu}{1 + (k-1)\mu} \sum_{i=1}^{k} z_i^2, \tag{47}$$

$$z_k^2 \geq \lambda_0 + \frac{(2k-3)\mu}{1 + (k-1)\mu} \sum_{i=1}^{k} z_i^2, \tag{48}$$

where $\Delta = n \cdot \mu$ in (47), then subset A_1 is the type-I optimal subset.

To prove the above theorem, we will show that for subsets having sizes equal to k, or sizes greater than k, or sizes less than k, the above three conditions will guarantee that subset A_1 is the type-I optimal subset. Since the proof is a little bit long and technical, it is postponed into Appendix B.4.

Remark 4.8 *Conditions (46), (47) and (48) are independent, i.e., none of them can be derived from the other two.*

The following theorem states the condition for set $A_1 = \{\phi_1, \phi_2, ..., \phi_k\}$ to be the type-II optimal subset.

Theorem 4.9 *Given λ and k, if*

$$\frac{\lambda}{2} - |z_{k+1}| \geq \frac{\sqrt{k}\mu}{1-(k-1)\mu}\sqrt{\sum_{i=1}^{k}\left(|z_i| + \frac{\lambda}{2}\right)^2}, \tag{49}$$

then subset A_1 is the type-II optimal subset.

Again, The proof is postponed into Appendix B.5 because it is a little bit technical.

The following corollary gives a sufficient condition for A_1 to be the concurrent optimal subset.

Corollary 4.10 *Given (46), (47), (48), and (49), subset A_1 is the concurrent optimal subset.*

6.4.2.1 Restrictiveness of the Aforementioned Sufficient Conditions

Readers may notice that the four conditions in the previous section are restrictive. One can easily find an example that does not satisfy these conditions, however still has the concurrent optimal subset A_1.

A counter example can be established as follows. Suppose n, m, and k are three positive integers satisfying $n > m > k$ and $n \geq m + k$. Let a_i denote the ith entry of vector $\mathbf{a} \in \mathbb{R}^k$ with $|a_1| \geq |a_2| \geq \cdots \geq |a_k|$. Let $I_{m\times m} \in \mathbb{R}^{m\times m}$ be an identity matrix and $\Phi_a \in \mathbb{R}^{k\times k}$ be the diagonal matrix with the ith diagonal entry being equal to a_i. Consider

$$\Phi = \text{standardized}\left\{\begin{pmatrix} \Phi_a \ \mathbf{0}_{k\times(m-k)} \\ I_{m\times m} \\ \mathbf{0}_{(n-k-m)\times m} \end{pmatrix}\right\}, \qquad y = \sum_{i=1}^{k}\phi_i,$$

where standardized$\{M\}$ refers to the standardization of all the columns of matrix M, matrices $\mathbf{0}_{k\times(m-k)}$ and $\mathbf{0}_{(n-k-m)\times m}$ are made by zeros, and ϕ_i is the ith column of Φ. The optimal solution is the first k covariates, and these covariates have larger correlations with y. However, there are many choices of m, n, k and vector $\mathbf{a}$, with which condition (46) is not satisfied. As a special case, consider the following simple example: $n = 10, m = 7, k = 3$, and $\mathbf{a} = (-1\ 1\ 0)^T$. It's not hard to verify that $\mu(\Phi) = 0.1667, z_3 = 0.7379, z_4 = -0.3162, [1-(k-1)\mu]z_k^2 = 0.3630$, and $2(k-1)^2\mu + z_{k+1}^2[1+(k-1)\mu] = 0.9117$. Hence, (46) does not hold for this case.

6.4.3 Sufficient Conditions based on the Model Matrix and the Correlations with Residuals

It is evident that the conditions in the previous subsection is restrictive. However, the derivation of the results (e.g., Theorem 4.7) demonstrates some key quantities that are required in the analysis: e.g., the correlations among the covariates, the correlations between the covariates and the response.

In order to come up with a practical subset selection scheme, it is helpful to have a sufficient condition for the type-I optimal subset. For example, when a solution path of **(P1)** is computed by an efficient stepwise algorithm, this sufficient condition can be used to test whether any of the solutions on this solution path is also type-I optimal. If yes, then a concurrent optimal subset is obtained.

We develop some sufficient conditions to identify whether a subset is a type-I optimal subset. Recall that $x \in \mathbb{R}^m$ denote a coefficient vector. Denote the corresponding residual vector by $\varepsilon = y - \Phi x$. Recall that $y \in \mathbb{R}^n$ and $\Phi \in \mathbb{R}^{n\times m}$ are the response vector and the model matrix, respectively. Let Ω denote the support of the vector x: $\Omega = \text{supp}(x)$. For an integer $k \geq 1$, let

$$\sigma^2_{\min,k} = \inf \frac{\|\Phi\delta\|_2^2}{\|\delta\|_2^2}, \quad \text{subject to } \|\delta\|_0 \leq k.$$

The above quantity reflects certain property of the model matrix. Furthermore, for a vector

$v \in \mathbb{R}^n$ and an integer $k \geq 1$, we define

$$c(v, k) = \sqrt{\sum_{i=1}^{k} v_{(i)}^2},$$

where $|v_{(1)}| \geq |v_{(2)}| \geq \cdots \geq |v_{(n)}|$ are the non-increasing-ordered magnitudes of the entries of vector v. For finite k, we assume that quantities $c^2(\Phi^T \varepsilon, k)$ and $\sigma^2_{\min,k}$ are available.

The following theorem provides a sufficient condition for a subset being included in a type-I optimal subset with respect to λ_0.

Theorem 4.11 *Given a subset of coefficient Ω. Suppose that coefficient vector x is the minimizer of function $\|y - \Phi x\|_2^2$ subject to supp$(x) \subset \Omega$. Let $\varepsilon = y - \Phi x$.*

(1) If $\min_{i \in \Omega} |x_i| > q_1(|\Omega|)$, then with respect to λ_0, there is no type-I optimal subset whose size of the support is less than $|\Omega|$.

(2) Furthermore, if $\min_{i \in \Omega} |x_i| > q(|\Omega|)$, then with respect to λ_0, we have $\Omega \subset \Omega'$, where Ω' is the type-I optimal subset with respect to λ_0.

The quantities $q_1(\cdot)$ and $q(\cdot)$ are defined as follows. For an integer $k \geq 1$,

$$\begin{aligned} q_1(k) &= \sup_{m<k} \frac{c(\Phi^T\varepsilon, 1) + \sqrt{c^2(\Phi^T\varepsilon, k+m) + (k-m)\lambda_0 \sigma^2_{\min,k+m}}}{\sigma^2_{\min,k+m}}, \\ q_2(k) &= \sup_{m \geq k} \frac{c(\Phi^T\varepsilon, 1) + \sqrt{c^2(\Phi^T\varepsilon, k+m) + (k-m)\lambda_0 \sigma^2_{\min,k+m}}}{\sigma^2_{\min,k+m}}, \end{aligned}$$

and

$$q(k) = \max\{q_1(k), q_2(k)\}.$$

Note that quantities $q_1(\cdot)$ and $q_2(\cdot)$ have the same objective function. However, the ranges of variable m are different. Because $q_1(k)$ only requires a finite choice of variable m, it is computable. It is not straightforward that for any $k \geq 1$, the quantity $q_2(k)$ exists. In this paper, we assume the existence of this quantity.

Proof. Suppose Ω' is the type-I optimal subset, with corresponding coefficient vector x'. We must have

$$\|y - \Phi x'\|_2^2 + \lambda_0 \|x'\|_0 \leq \|y - \Phi x\|_2^2 + \lambda_0 \|x\|_0. \tag{50}$$

Denote $\delta = x' - x$, we have $\|\delta\|_0 \leq |\Omega| + |\Omega'|$. We will prove that

$$\text{“if } |\Omega'| < |\Omega|, \text{ then } \|\delta\|_\infty \leq q_1(\Omega),\text{”} \tag{51}$$

and

$$\text{“for any } \Omega', \|\delta\|_\infty \leq q(\Omega).\text{”} \tag{52}$$

To see the above, a reformulation of (50) gives

$$\|\varepsilon - \Phi\delta\|_2^2 \leq \|\varepsilon\|_2^2 + \lambda_0(|\Omega| - |\Omega'|),$$

which is equivalent to

$$\|\Phi\delta\|_2^2 \leq 2\langle \Phi^T\varepsilon, \delta\rangle + \lambda_0(|\Omega| - |\Omega'|), \tag{53}$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product between two sequences. Define $\delta' = \sigma^2_{\min,|\Omega|+|\Omega'|} \cdot \delta$. Because $\|\Phi\delta\|_2^2 \geq \sigma^2_{\min,|\Omega|+|\Omega'|}\|\delta\|_2^2$, and (53), we have

$$\|\delta'\|_2^2 \leq 2\langle \Phi^T\varepsilon, \delta'\rangle + \lambda_0(|\Omega| - |\Omega'|) \cdot \sigma^2_{\min,|\Omega|+|\Omega'|}.$$

The above is equivalent to

$$\|\Phi^T\varepsilon - \delta'\|_2^2 \leq \|\Phi^T\varepsilon\|_2^2 + \lambda_0(|\Omega| - |\Omega'|) \cdot \sigma^2_{\min,|\Omega|+|\Omega'|}.$$

Define $\varepsilon^* = \Phi^T\varepsilon$. The above inequality leads to

$$\sum_{i \in \Omega \cup \Omega'} (\varepsilon_i^* - \delta_i')^2 \leq \sum_{i \in \Omega \cup \Omega'} (\varepsilon_i^*)^2 + \lambda_0(|\Omega| - |\Omega'|) \cdot \sigma^2_{\min,|\Omega|+|\Omega'|}.$$

The above immediately leads to

$$\sup_{i \in \Omega \cup \Omega'} |\delta_i'| \leq \sup_{i \in \Omega \cup \Omega'} |\varepsilon_i^*| + \sqrt{\sum_{i \in \Omega \cup \Omega'} (\varepsilon_i^*)^2 + \lambda_0(|\Omega| - |\Omega'|) \cdot \sigma^2_{\min,|\Omega|+|\Omega'|}}.$$

Dividing both sides by $\sigma^2_{\min,|\Omega|+|\Omega'|}$, we have

$$\sup_{i \in \Omega \cup \Omega'} |\delta_i| \leq \frac{c(\Phi^T\varepsilon, 1) + \sqrt{c^2(\Phi^T\varepsilon, |\Omega| + |\Omega'|) + \lambda_0(|\Omega| - |\Omega'|) \cdot \sigma^2_{\min,|\Omega|+|\Omega'|}}}{\sigma^2_{\min,|\Omega|+|\Omega'|}}. \tag{54}$$

Recall the definitions of $q_1(\cdot)$ and $q(\cdot)$, (51) and (52) can be derived directly from (54).

Now we are able to verify item (1) in the theorem. Suppose there is a type-I optimal subset Ω' satisfying $|\Omega'| < |\Omega|$. We have

$$|x_i'| \geq |x_i| - |x_i - x_i'| \geq |x_i| - q_1(\Omega) > 0.$$

The second inequality is based on (51); and the last inequality is from the condition in item (1). The above implies $\Omega \subset \Omega'$, which contradicts $|\Omega'| < |\Omega|$. We have proved item (1).

The proof of item (2) is strongly similar to the proof of (1). We skip the obvious details.

□

The above theorem is motivated by a recent related work in applied mathematics. Readers may compare it with the test proposed in [36]. Their test is related to the optimality in sparse signal representations.

In Theorem 4.11, quantities $q_1(\cdot)$ and $q(\cdot)$ require multiple values of $\sigma^2_{\min,k}$, for a range of values of k. Comparing to the quantities $c(\cdot, k)$, it is harder to compute $\sigma^2_{\min,k}$'s. Inspired by the derivation in Theorem 2 in [36], we derive a sufficient condition, which only depends on $\sigma^2_{\min,|\Omega|}$, where Ω is the subset that is tested. To state our result, the following quantity needs to be defined: for an integer $m \geq 1$ and a given integral constant M, let

$$\lambda(m; M) = 1 - \frac{M}{\sqrt{m}} \sup_{|\mathcal{I}| \leq m} \sup_{k \notin \mathcal{I}} \|\Phi_{\mathcal{I}}^{+} \phi_k\|_2,$$

where $\mathcal{I}$ is a subset of indices, $|\mathcal{I}|$ denotes the size of this subset, matrix $\Phi_{\mathcal{I}}$ is a submatrix of Φ whose column indices form the set $\mathcal{I}$, $\Phi_{\mathcal{I}}^{+} = (\Phi_{\mathcal{I}}^{*}\Phi_{\mathcal{I}})^{-1}\Phi_{\mathcal{I}}^{*}$ is the Moore-Penrose pseudo-inverse [35] with $(\cdot)^*$ denoting the adjoint, and ϕ_k is the kth column (i.e., covariate) in Φ. Given m, quantity $\lambda(m)$ can be computed by enumerating all m-subset of the covariates.

Now we present another sufficient condition.

Theorem 4.12 *Given a subset of coefficient Ω. Suppose that coefficient vector x is the minimizer of function $\|y - \Phi x\|_2^2$ subject to supp$(x) \subset \Omega$. Suppose it is known a priori that the size of the type-I optimal subset is no larger than M. If $\min_i |x_i| > q'(|\Omega|, M)$, then set Ω is at least a subset of the type-I optimal subset. Here quantity $q'(\cdot)$ is defined as, for integer $k \geq 1$ and constant M,*

$$q'(k, M) = \sup_{1 \leq m \leq M} \frac{c(\Phi^T\varepsilon, 1) + \sqrt{c^2(\Phi^T\varepsilon, k) + \lambda_0 \cdot \frac{k^2(k-m)}{(k+m)^2} \cdot \sigma^2_{\min,k} \cdot \lambda^2(k;m)}}{\frac{k}{k+m}\sigma^2_{\min,k} \cdot \lambda^2(k;m)}$$

Proof. The beginning of the proof is the same as the proof of the previous theorem. It starts to deviates at stage (53). For readers' convenience, we restate the inequality (53):

$$\|\Phi\delta\|_2^2 \leq 2\langle \Phi^T\varepsilon, \delta\rangle + \lambda_0(|\Omega| - |\Omega'|). \tag{55}$$

Readers are referred to the previous proof for the meanings of the notations.

First, we have

$$\langle \Phi^T \varepsilon, \delta \rangle \leq \sum_{i=1}^{n} |b_{(i)}| \cdot |\delta_{(i)}|, \tag{56}$$

where $|\delta_{(1)}| \geq |\delta_{(2)}| \geq \cdots \geq |\delta_{(n)}|$ is the ordered list of the magnitudes of the entries in vector δ. Similarly, $|b_{(1)}| \geq |b_{(2)}| \geq \cdots \geq |b_{(n)}|$ is the ordered list of the magnitudes of the entries in vector $\Phi^T \varepsilon$. We denote $\Phi^T \varepsilon$ by b. The following manipulations are needed:

$$\begin{aligned}
\text{R.H.S. of (56)} &= \sum_{i=1}^{|\Omega|} |b_{(i)}| \cdot |\delta_{(i)}| + \sum_{i=|\Omega|+1}^{n} |b_{(i)}| \cdot |\delta_{(i)}| \\
&\leq \sum_{i=1}^{|\Omega|} |b_{(i)}| \cdot |\delta_{(i)}| + |b_{(|\Omega|+1)}| \cdot \sum_{i=|\Omega|+1}^{n} |\delta_{(i)}| \\
&\leq \sum_{i=1}^{|\Omega|} |b_{(i)}| \cdot |\delta_{(i)}| + |b_{(|\Omega|+1)}| \cdot \frac{|\Omega'|}{|\Omega|} \cdot \sum_{i=1}^{|\Omega|} |\delta_{(i)}| \\
&\leq \left(1 + \frac{|\Omega'|}{|\Omega|}\right) \sum_{i=1}^{|\Omega|} |b_{(i)}| \cdot |\delta_{(i)}| \\
&= \left(1 + \frac{|\Omega'|}{|\Omega|}\right) \langle b^*, \delta^*_{|\Omega|} \rangle,
\end{aligned} \tag{57}$$

where vector $\delta^*_{|\Omega|}$ takes the absolute values of δ only at the positions where vector δ has the $|\Omega|$ largest magnitudes and zeros elsewhere. I.e.,

$$\delta^*_{|\Omega|,i} = \begin{cases} |\delta_i|, & \text{if } |\delta_i| \geq |\delta_{(|\Omega|)}|; \\ 0, & \text{elsewise.} \end{cases}$$

For vector b^*,

$$b^*_i = |b_{(j)}|, \quad \text{where } \delta_i = \delta_{(j)}.$$

Putting (56) and (57) together, we have

$$\langle \Phi^T \varepsilon, \delta \rangle \leq \left(1 + \frac{|\Omega'|}{|\Omega|}\right) \langle b^*, \delta^*_{|\Omega|} \rangle. \tag{58}$$

Meanwhile, for any Ω, we have

$$\begin{aligned}
\|\Phi\delta\|_2^2 &\geq \|\Phi_\Omega \Phi_\Omega^+ \Phi\delta\|_2^2 \\
&\geq \sigma^2_{\min,|\Omega|} \cdot \|\Phi_\Omega^+ \Phi\delta\|_2^2 \\
&= \sigma^2_{\min,|\Omega|} \cdot \|\delta_\Omega + \Phi_\Omega^+ \Phi_{\Omega^c} \delta_{\Omega^c}\|_2^2,
\end{aligned} \tag{59}$$

where set Ω^c is the complement of set Ω, matrices Φ_Ω and Φ_{Ω^c} are submatrices of matrix Φ by taking columns whose indices are in Ω and Ω^c, respectively. As mentioned earlier, matrix Φ_Ω^+ is a pseudo-inverse of Φ_Ω. Vector δ_Ω (respectively, δ_{Ω^c}) only takes nonzero values when the index is in the set Ω (respectively, Ω^c). Note here Ω can be any subset of the indices, which is different with the Ω in the assumption at the beginning of the proof – we have an abuse of the notation. In the above steps, the first inequality is true because the matrix $\Phi_\Omega \Phi_\Omega^+$ is a projection matrix. The second inequality is based on the definition of $\sigma^2_{\min,|\Omega|}$. The last step is just a reorganization. Furthermore, we have

$$\begin{aligned}
\|\delta_\Omega + \Phi_\Omega^+ \Phi_{\Omega^c}\delta_{\Omega^c}\|_2 &\geq \|\delta_\Omega\|_2 - \sum_{k\in\Omega^c} |\delta_k| \cdot \sup_{k\notin\Omega} \|\Phi_\Omega^+ \phi_k\|_2 \\
&\geq \|\delta_\Omega\|_2 - \sum_{k=|\Omega|+1}^{n} |\delta_{(k)}| \cdot \sup_{k\notin\Omega} \|\Phi_\Omega^+ \phi_k\|_2 \\
&\geq \|\delta_\Omega\|_2 - \frac{|\Omega'|}{|\Omega|} \cdot \|\delta^*_{|\Omega|}\|_1 \cdot \sup_{k\notin\Omega} \|\Phi_\Omega^+ \phi_k\|_2 \\
&\geq \|\delta_\Omega\|_2 - \frac{|\Omega'|}{\sqrt{|\Omega|}} \cdot \|\delta^*_{|\Omega|}\|_2 \cdot \sup_{k\notin\Omega} \|\Phi_\Omega^+ \phi_k\|_2 \\
&\geq \lambda(|\Omega|; |\Omega'|) \cdot \|\delta^*_{|\Omega|}\|_2. \qquad (60)
\end{aligned}$$

In the above, the first and the second steps are common maneuvers. The third inequality is based on $\|\delta^*_{|\Omega|}\|_1/|\Omega| \geq \sum_{k=|\Omega|+1}^n |\delta_{(k)}|/|\Omega'|$. The fourth inequality is based on $\|\delta^*_{|\Omega|}\|_1 \leq \sqrt{|\Omega|} \cdot \|\delta^*_{|\Omega|}\|_2$. The last step recalls the definition of $\lambda(\cdot,\cdot)$ and takes Ω as the indices subset where $\delta^*_{\|\Omega\|}$ having nonzero entries. Combining (59) and (60), we have

$$\|\Phi\delta\|_2^2 \geq \sigma^2_{\min,|\Omega|} \cdot \lambda^2(|\Omega|; |\Omega'|) \cdot \|\delta^*_{|\Omega|}\|_2^2. \qquad (61)$$

Now we put the above results together, and then maneuver back to the argument as in the proof of Theorem 4.11. Combining (55), (58), and (61), we have

$$\sigma^2_{\min,|\Omega|} \cdot \lambda^2(|\Omega|; |\Omega'|) \cdot \|\delta^*_{|\Omega|}\|_2^2 \leq 2\left(1 + \frac{|\Omega'|}{|\Omega|}\right) \langle b^*, \delta^*_{|\Omega|}\rangle + \lambda_0(|\Omega| - |\Omega'|).$$

Let

$$\delta' = \frac{|\Omega|}{|\Omega| + |\Omega'|}\sigma^2_{\min,|\Omega|} \cdot \lambda^2(|\Omega|; |\Omega'|) \cdot \delta^*_{|\Omega|}.$$

We have

$$\|\delta'\|_2^2 \leq 2\langle b^*, \delta'\rangle + \lambda_0 \cdot \frac{|\Omega|^2(|\Omega| - |\Omega'|)}{(|\Omega| + |\Omega'|)^2} \cdot \sigma^2_{\min,|\Omega|} \cdot \lambda^2(|\Omega|; |\Omega'|).$$

The above is equivalent to

$$\|\delta' - b^*\|_2^2 \leq \|b^*\|_2^2 + \lambda_0 \cdot \frac{|\Omega|^2(|\Omega| - |\Omega'|)}{(|\Omega| + |\Omega'|)^2} \cdot \sigma^2_{\min,|\Omega|} \cdot \lambda^2(|\Omega|; |\Omega'|).$$

The above leads to the following

$$\|\delta'\|_\infty \leq \|b^*\|_\infty + \sqrt{c^2(b^*, |\Omega|) + \lambda_0 \cdot \frac{|\Omega|^2(|\Omega| - |\Omega'|)}{(|\Omega| + |\Omega'|)^2} \cdot \sigma^2_{\min,|\Omega|} \cdot \lambda^2(|\Omega|; |\Omega'|)}.$$

Recall the definition of δ' and b^*, we have

$$\begin{aligned} \|\delta\|_\infty &\leq \|b\|_\infty + \sqrt{c^2(b, |\Omega|) + \lambda_0 \cdot \frac{|\Omega|^2(|\Omega| - |\Omega'|)}{(|\Omega| + |\Omega'|)^2} \cdot \sigma^2_{\min,|\Omega|} \cdot \lambda^2(|\Omega|; |\Omega'|)} \\ &\leq \frac{c(\Phi^T\varepsilon, 1) + \sqrt{c^2(\Phi^T\varepsilon, |\Omega|) + \lambda_0 \cdot \frac{|\Omega|^2(|\Omega| - |\Omega'|)}{(|\Omega| + |\Omega'|)^2} \cdot \sigma^2_{\min,|\Omega|} \cdot \lambda^2(|\Omega|; |\Omega'|)}}{\frac{|\Omega|}{|\Omega| + |\Omega'|}\sigma^2_{\min,|\Omega|} \cdot \lambda^2(|\Omega|; |\Omega'|)} \\ &\leq q'(|\Omega|; M). \end{aligned} \tag{62}$$

The above is equivalent to $\|x - x'\|_\infty < q'(|\Omega|; M)$. Using the same argument as in the last proof, we can argue that $\Omega \subset \Omega'$. Suppose $x_i \neq 0$, we have

$$|x_i'| \geq |x_i| - |x_i - x_i'| \geq |x_i| - q'(|\Omega|, M) > 0,$$

which implies that $\Omega \subset \Omega'$. □

6.4.3.1 Application in the Case with Orthonormal Covariates

If the model matrix Φ is orthonormal, readers can verify that $\sigma^2_{\min,k} = 1$ and $\lambda(m; M) = 1$. It brings significantly simplified criteria in Theorem 4.11 and Theorem 4.12. Comparing with the result in Theorem 3.5, the new criteria are less attractive. We consider this a price of the generality.

It will be interesting to apply the above conditions to some applications with real data sets. However, due to the length of this paper, considering we are more focused on the formulation and theoretical developments in the present paper, we leave applications for future publications.

6.5 Discussion

6.5.1 Computing Versus Statistical Properties

The question that we addressed in this paper is quite different from some statistical works. In the present paper, we identify easy to verify (polynomial time) conditions for the type-I optimal subset. Our direct motivation is that certain greedy algorithm can find a path of type-II optimal subsets. If one of these type-II optimal subset is confirmed to be type-I optimal, then a concurrent optimal subset is obtained. In the above sense, our question is more statistical computing than prediction.

In traditional approaches of subset selection, researchers try to answer the questions regarding the consistency of variable selection, as well as the optimal accuracy rate in submodel prediction. There is a large scope of existing efforts. It is impossible and unnecessary for us to give a comprehensive survey here. We will just list some publications that have been informative and inspiring to us. [25], [95], [24], [85], [103], and the references therein give some interesting results in model estimation integrating the prediction accuracy. Consistency of variable selection has been studied in [101].

Nowadays, due to the rapid rising of data sizes, it becomes increasingly important to develop computationally efficient statistical principle. Our idea of finding efficient sufficient conditions for otherwise unsolvable (i.e., NP-hard) subset selection principle is an incarnation of the aforementioned ideology.

6.5.2 Other Works in Variable Selection

Despite their generality, the formulations of **(P0)** and **(P1)** do not cover all the existing works in statistical model selection. We review some recent works that have attracted our attention.

Paper [27] proposes a family of new variable selection methods based on a nonconcave penalized likelihood approach. The criterion is to minimize

$$\text{Fan\&Li} = RSS(x) + 2n \cdot \sum_{j=1}^{\|x\|_0} p_\lambda(|\theta_j|),$$

where $p_\lambda(\cdot)$ is a penalty function which is symmetric, nonconcave on $(0, \infty)$ and has singularities at origin. With proper choice of λ, Fan and Li show that the estimators would have good statistical properties, such as sparsity and asymptotic normality.

Shen and Ye in [86] suggest an adaptive model selection procedure to estimate the algorithmic parameter λ from the data. In detail, the optimal value of λ is obtained by minimizing

$$\text{Shen\&Ye} = RSS(x) + \hat{g}_0(\lambda_0) \cdot \sigma^2,$$

which is derived from the optimal estimator of the loss $l(\theta, \hat{\theta})$. Quantity $\hat{g}_0(\lambda_0)$ is the estimator of $g_0(\lambda_0)$, which is independent of the unknown parameter θ. Value $g_0(\lambda_0)/2$ is called the generalized degrees of freedom in [98].

At this moment, we do not know whether there are analogous conditions (to those in Section 6.4.3) that can be established in the above two settings. Examining possible connections will be an interesting topic for future research.

6.5.3 Back Elimination

Subset selections include at least three basic approaches: forward selection, backward elimination, and all subset selection. Problem **(P0)** is an all subset selection method. The greedy algorithms that have been discussed in this paper are assumed to be forward selection algorithms. Readers are referred to Section 6.2.2.

In [15], a very interesting result is proved for backward elimination. It is shown that under certain conditions, back elimination finds the solution of **(P0)**. Such a result reveals the properties of problem **(P0)** from another angle.

It will be interesting to examine whether the approaches that are adopted in Section 6.4.3 can lead to stronger conditions in back elimination approaches. Again, this is left as a topic of future research.

6.5.4 Other Greedy Algorithms and Absolutely Optimal Subset in Variable Selection

We have treated LARS as a forward stepwise algorithm. Other greedy algorithms have made significant impact in other fields, such as signal processing. Two representative ones are matching pursuit (MP) [16, 64] and an improved version – orthogonal matching pursuit (OMP) [79]. MP and OMP do not generate the regularized solution path, while a version of LARS does. However, the intensive research effort following MP and OMP will provide researchers powerful tools.

Researchers have studied on the subsets that are unconditionally concurrent optimal, i.e., its concurrent optimality depends on neither the coefficients nor the corresponding residuals. The representative works include [19], [91], and [92]. The concept of exact recovery coefficient (ERC) [92] has inspired many recent works. Readers can compare ERC with our quantity $\lambda(m; M)$ that is defined right before Theorem 4.12 in Section 6.4.3.

Note that in our sufficient conditions, both coefficient and residuals are taken into account. This is due to the different emphasis of the problem. Comparing with our works, the results mentioned in the last paragraph can be considered as analysis of the worst cases.

6.5.5 Other Related Topics

An interesting model selection approach that adopts Bayesian computing is presented in [14]. This provides another interesting aspect of strategies. It will be interesting to analyze the connection with the contents of this paper.

Variable selection is a critical problem in supersaturated design. A citation search of [96] will provide most of existing literature. A numerically efficient condition on the optimality of subsets has the potential to identify a good design. Further study of this problem is left as a topic of future research.

6.6 Conclusion

Stepwise algorithms can be numerically efficient, i.e., polynomial time. Specially designed stepwise algorithms can find type-II optimal subset in subset selection. We derived sufficient conditions to test whether these type-II optimal subsets are also type-I optimal. Such an approach renders polynomial time algorithms to locate concurrent optimal subsets, which otherwise requires solving an NP-hard optimization problem in general.

CHAPTER VII

REGRESSIONS BY ENHANCED LEAPS-AND-BOUNDS VIA ADDITIONAL OPTIMALITY TESTS (LBOT)

The conditions derived in the last chapter is valuable. In this chapter, we extend the results into the implementation of certain all-subset selection algorithm. In exhaustive subset selection in regressions, the leaps-and-bounds algorithm by Furnival and Wilson [28] is the current state-of-the-art. It utilizes a branch and bound strategy. We improve it by introducing newly designed optimality tests, retaining the original general framework. Compared with the original leaps-and-bounds algorithm, the proposed method further reduces the number of subsets that are needed to be considered in the exhaustive subset search. Simulations demonstrate the improvements in numerical performance. Our new description of the leaps-and-bounds algorithm, which is based on our newly designed *pair tree*, is independent of programming languages, and therefore is more accessible.

This chapter is organized as follows. In Section 7.1, we state our objective, bring in the leaps-and-bounds algorithm, and summarized our contributions. In Section 7.2, some basic results regarding the fast computation of RSS's and matrix inverse are given. In Section 7.3, a specific version of the LB method in [28] is reviewed, and will serve as a starting point of our algorithmic description. In Section 7.4, additional optimality tests are derived. In Section 7.5, the newly derived optimality tests are integrated with LB, and the new leaps-and-bounds method (i.e., LBOT) is established. In Section 7.6, simulations are provided to demonstrate the improvements of performance. Some discussions and conclusions are provided in sections 7.7 and 7.8, respectively.

7.1 Introduction

We continue studying the variable selection problem in a generic regression model in this chapter. Again, regression model can be expressed as follows:

$$y = \Phi x + \varepsilon,$$

where $y \in \mathbb{R}^n$ is a response vector, n is the number of observations, matrix $\Phi \in \mathbb{R}^{n\times(m+1)}$, $\Phi = [\mathbf{1}_n/\sqrt{n}, \phi_1, \ldots, \phi_m]$, is the model matrix with a constant column $\phi_0 = \mathbf{1}_n/\sqrt{n}$ and covariates $\phi_i \in \mathbb{R}^n, 1 \leq i \leq m$, and vector $\varepsilon \in \mathbb{R}^n$ is a random vector. The model selection is to choose a subset of $\{\phi_i : i = 1, 2, \ldots, m\}$, so that the regression model based on the selected subset is as effective in prediction as the model built on the full set of covariates. There is a huge related literature in statistics, e.g., model estimation theory, which is not the theme of this dissertation. We will concentrate on the leaps-and-bound (LB) algorithm [28], which is a widely used subset selection method based on all-subsets comparisons. Recent papers – [55] and [29] – give excellent surveys on *subset selection*.

In [28], the following problem is solved: for all integer k, $1 \leq k \leq m$,

$$\begin{array}{lll} \text{(FW)} & \min_x & \|y - \Phi x\|_2^2, \\ & \text{subject to:} & \|x\|_0 = k, \end{array}$$

where $\|\cdot\|_2^2$ denotes the sum of squares of the elements in a vector (i.e., the square of the ℓ_2 vector norm), and $\|\cdot\|_0$ is the number of nonzero entries in a vector (which is also called ℓ_0 quasi-norm). We name the problem (FW) to recognize the contribution of the original proposers of LB. Solving (FW) gives a way to realize model selection. It is connected with many widely used model selection methods, which can be summarized as the following optimization problem:

$$\min_x \quad \|y - \Phi x\|_2^2 + \lambda_0 \cdot \|x\|_0, \tag{63}$$

where λ_0 is an algorithmic parameter. For AIC and C_p, we have $\lambda_0 = 2\widehat{\sigma}^2$, where $\widehat{\sigma}^2$ is an unbiased estimate of the common variance of the random errors. For BIC and MDL, we have $\lambda_0 = \sigma^2 \log n$. We refer to [53] for more relevant information. The foregoing paper also proves that problem (63) is NP-hard. Readers may compare the difference between (63) and (FW). Note that solutions to (FW) lead to a solution to (63), with a small amount of additional computation.

Since the initial introduction of LB, little effort has been reported to improve this algorithm. The essence of the LB method is a branch and bound procedure, which uses tests to reduce the number of subsets that should be considered in an exhaustive subset search. In this dissertation, in the same branch-and-bound framework, we derive new optimality tests. It is shown that the induced additional tests can further reduce the number of subsets that are required to be considered. Hence, it accelerates LB. The derived method is named *leaps-and-bounds via optimality tests* (LBOT).

We briefly describe the motivation for the new tests. The original LB algorithm utilizes the following optimality test. Let A and B denote two distinct subsets of covariates, and assume that A is a subset of B: $A \subset B$. Let RSS(A) (resp., RSS(B)) denote the residual sum of squares of the regression model that is built on subset A (resp., B). We have RSS(A) $\geq$ RSS(B). In this thesis, more powerful optimality tests will be derived. The key idea is deriving a more strict necessary condition for a subset to outperform an existing optimal subset. *We will not only use the residual sums of squares, which are utilized in the LB, but also consider the coefficients and the residuals associated with the optimal subsets.* Details regarding the derivation of such a condition are presented in Section 7.4.2.

Simulations demonstrate the improvement in performance. They also indicate the situations in which LB and its enhanced one – LBOT – are likely to significantly reduce the number of subsets that are needed to be computed. We will argue in this chapter that the number of subsets that are examined is a good indicator of the computational complexity, because of its implementational independence. Some heuristics that can possibly improve the performance of our algorithms are tested, and the results are presented.

7.2 Review of Basics

Some relevant computational details are presented here. The ideas can be found in the original paper [28]. The purposes of re-presenting them are

- to demonstrate that from one subset to a new subset, by inserting or deleting a covariate, there is an efficient numerical approach for the computation regarding the submodels;

- to make a point that the number of subsets that are needed to be computed in an algorithm (e.g., LB or LBOT) is an indicator of the complexity of this algorithm.

7.2.1 Computing Regarding Submodels

From now on, for simplicity, we assume that the covariates $\phi_i, i = 1, 2, \ldots, m$, are standardized, i.e., for $1 \leq i \leq m$, $\text{ave}(\phi_i) = 0$, and $\|\phi_i\|_2 = 1$, where $\text{ave}(\cdot)$ (resp., $\|\cdot\|_2$) denotes the average (resp., ℓ_2-norm) of a vector. It is evident that the correlation matrix among the response and the covariates is

$$(y, \Phi)^T(y, \Phi) = \begin{pmatrix} y^T y & y^T \Phi \\ \Phi^T y & \Phi^T \Phi \end{pmatrix}.$$

Moreover, the diagonal entries of matrix $\Phi^T \Phi$ are all equal to 1.

A submodel is determined by a subset of the covariates. Suppose subset Ω ($\Omega \subset \{1, 2, \ldots, m\}$) determines the submodel. The corresponding model matrix is made by the constant column vector and the columns having indices from Ω. The model matrix is denoted by Φ_Ω: $\Phi_\Omega = [\mathbf{1}_n/\sqrt{n}, \{\phi_i\}_{i\in\Omega}]$. Note the first column corresponds to the intercept term, which is also included in submodels. Let $\widehat{\beta}(\Omega)$ denote the least square fit on this submodel, we have

$$\widehat{\beta}(\Omega) = (\Phi_\Omega^T \Phi_\Omega)^{-1} \Phi_\Omega^T \cdot y. \tag{64}$$

Let $\text{RSS}(\Omega)$ denote the residual sum of squares of the least square fit, we have

$$\text{RSS}(\Omega) = y^T y - (\Phi_\Omega^T \cdot y)^T (\Phi_\Omega^T \Phi_\Omega)^{-1} (\Phi_\Omega^T \cdot y). \tag{65}$$

Note $\Phi_\Omega^T \cdot y$ is a subvector of $\Phi^T \cdot y$, which can be handily read from the correlation matrix.

7.2.2 Two Basic Linear Algebra Results

The following simple linear algebra results show that when adding or deleting one covariate, the resulting inverse matrix can be computed efficiently. The original LB paper also took advantage of these facts. However, their description is less direct.

Lemma 2.1 *For a positive integer k, given symmetric matrix $M \in \mathbb{R}^{k \times k}$ with its inverse M^{-1}, a vector $v \in \mathbb{R}^k$, and a constant $c \in \mathbb{R}$, we have*

$$\begin{pmatrix} M & v \\ v^T & c \end{pmatrix}^{-1} = \begin{pmatrix} M^{-1} + \tau M^{-1} v v^T M^{-1} & -\tau M^{-1} v \\ -\tau v^T M^{-1} & \tau \end{pmatrix},$$

where scalar $\tau = (c - v^T M^{-1} v)^{-1}$.

The following is an easy extension.

Corollary 2.2 *Given the same notations as in Lemma 2.1, if we have*

$$\begin{pmatrix} M & v \\ v^T & c \end{pmatrix}^{-1} = \begin{pmatrix} B_{11} & B_{12} \\ B_{12}^T & B_{22} \end{pmatrix},$$

where $B_{11} \in \mathbb{R}^{k \times k}$, $B_{12} \in \mathbb{R}^{k \times 1}$, and $B_{22} \in \mathbb{R}$, then the following holds:

$$M^{-1} = B_{11} - B_{12} B_{12}^T / B_{22}.$$

Let $I(\Omega)$ denote the inverse matrix $(\Phi_\Omega^T \Phi_\Omega)^{-1}$. For $i \in \Omega$ and $j \notin \Omega$, from Lemma 2.1 (resp., Corollary 2.2), there is a fast way to compute $I(\Omega \cup \{j\})$ (resp., $I(\Omega \backslash \{i\})$). The following is in the original LB paper [28]. Readers can easily verify it.

Lemma 2.3 *For above mentioned indices i and j, subset Ω, and integer $k = |\Omega|$, which is the size of subset Ω, it takes $O(k^2)$ numerical operations (additions, subtractions, multiplications, and divisions) to generate the inverse matrices corresponding to adding/deleting one covariate to/from the subset Ω.*

7.3 Subset Arrangement and the Leaps-and-Bounds Algorithm

The ingenious idea in the original LB paper [28] is to introduce a systematic way to scan through all the subsets, at the same time, 'leaping' over those evidently nonoptimal subsets. Here, we redescribe their scheme. The pairing structure and the pair tree in Section 7.3.2 are new, which is motivated by the description in [28]. Because the original description requires knowledge in Fortran language and two trees, we believe our description is more understandable.

7.3.1 Inverse Tree

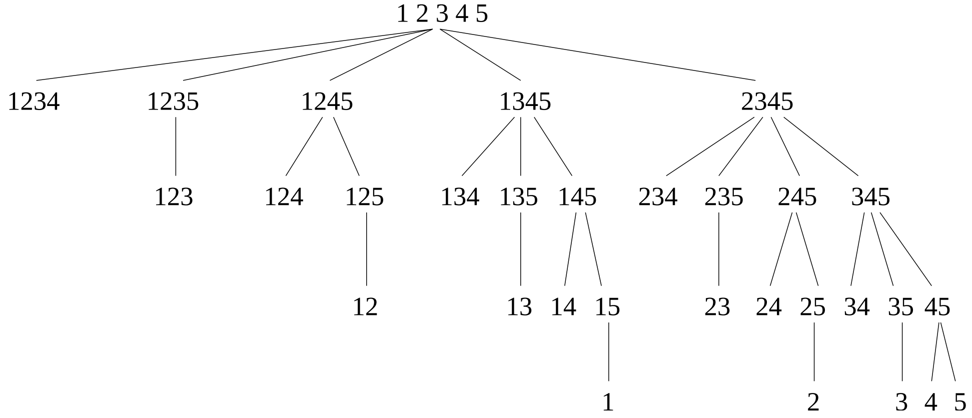


Figure 56: An inverse tree with $m = 5$. For simplicity, in figures, "1234" is equivalent with subset $\{1234\}$ in the text.

Figure 56 gives an inverse tree with $m = 5$ covariates. Its construction is in the original LB paper. For completeness, we briefly describe it in the following.

1. The root node is the full set $\{1, 2, \ldots, m\}$.
2. Level 1 is made by m ordered children of the root node by removing one covariate at a time from the full set at the decreasing order: $m, m-1, \ldots, 2, 1$.
3. Consider a node associated with subset $\{i_1 i_2 \cdots i_k\}, k \geq 1, i_1 < i_2 < \cdots < i_k$. Assume it is the jth ($j \geq 1$) child of its parent. At the next level, this node has $j - 1$ children that are generated by deleting one covariate at a time from the set $\{i_1 i_2 \cdots i_k\}$ with the order $i_k, i_{k-1}, \ldots, i_{k+2-j}$.
4. The tree stops growing when it reaches the subsets made by one covariate, or all terminal nodes are the first children.

Readers can easily verify the following facts, which are collectively presented in a theorem.

Theorem 3.1 *The inverse tree has the following properties:*

- *The above constructed inverse tree contains all the* $2^m - 1$ *subsets of* $\{1, 2, \ldots, m\}$.

- *Each subset appears once and only once in this tree.*

- *The sizes of the subsets associated with the nodes at level k $(1 \leq k \leq m-1)$ of this tree are equal to $m-k$.*

- *Each subset associated with a node in this tree, except the root node, is obtainable by removing one covariate from the subset associated with its parent node.*

The following observation will be utilized in a new description of the LB algorithm.

Theorem 3.2 *In the inverse tree with m covariates, the subtree rooted at node $\{2, 3, \ldots, m\}$ has the identical structure with the subtree started at the original root node, after pruning the subtree rooted at node $\{2, 3, \ldots, m\}$ and ignoring the only terminal nodes at the bottom level: $\{1\}$. Moreover, if Ω is a subset in the subtree rooted at node $\{2, 3, \ldots, m\}$, then subset $\Omega \cup \{1\}$ is associated with the node at the same position in the latter pruned tree.*

7.3.2 Pair Tree

The original description of the LB method is based on two trees: *regression* tree and *bound* tree. Being inspired by these two trees, we construct the following pairing scheme and a new tree for the pairs of subsets, so that the subsets searching and leaping can be realized based on one single structure. We believe this new scheme gives a more intuitive description. Figure 57 gives such a pair tree for the same case ($m = 5$) as depicted in Figure 56.

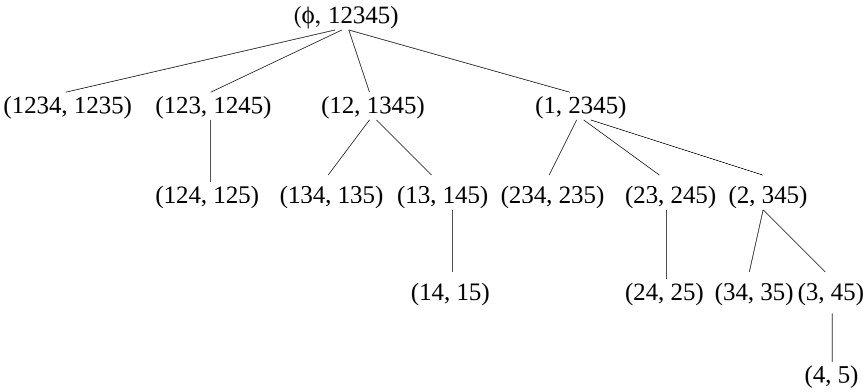
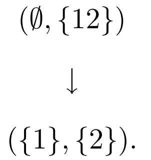


Figure 57: A pair tree with $m = 5$.

A pair tree can be constructed by using induction. Readers can use Theorem 3.2 to verify that the following construction is well defined. For $m = 2$, the pair tree (denoted as

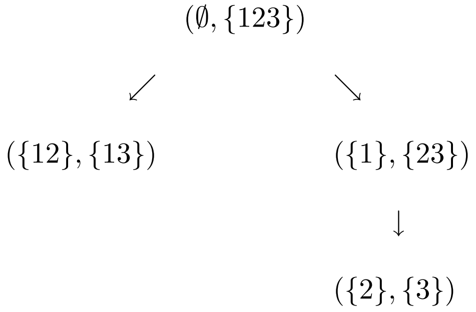
$PT(2)$) is



For $m = 3$, the corresponding pair tree is constructed by the following three steps:

- Transfer index i ($i = 1, 2$) in tree $PT(2)$ to $i + 1$. The generated tree is called T_1.
- Take a T_1, insert $\{1\}$ in all the nonempty subsets. The new tree is called T_2.
- Take another T_1, convert $\emptyset$ to $\{1\}$, and make it an additional subtree of T_2 by making the root node of the modified T_1 a new child of the root node of T_2. The new child is the last child at the first level of T_2. The combined pair tree is $PT(3)$.

The following depicts $PT(3)$:



In general, given $PT(m)$, $PT(m + 1)$ is generated by three steps: (1) Transfer index i ($i = 1, \ldots, m$) in tree $PT(m)$ to $i+1$. The generated tree is called T_1. (2) Take a T_1, insert $\{1\}$ in all the nonempty subsets. The new tree is called T_2. (3) Take another T_1, convert $\emptyset$ to $\{1\}$, make it an additional subtree of T_2 by making the root node of the modified T_1 a new child of the root node of T_2. The result pair tree is $PT(m + 1)$. Readers can observe the strong parallelism to the previous description. In fact, it is a generalization.

We can easily verify the following.

Theorem 3.3 *Consider a pair tree for m covariates (i.e., $PT(m)$).*

1. *In the aforementioned pair tree, each subset of $\{1, 2, \ldots, m\}$ appears once and only once.*

2. *For an intermediate node* (Ω_1, Ω_2)*, all the subsets in the descendant nodes are the subsets of* Ω_2*. This indicates that the order in a pair can not be changed.*

3. *For integers* $1 \leq k_1, k_2 \leq m$*, suppose two subsets in a pair contain* k_1 *and* k_2 *covariates, respectively. The sizes of the subsets in the descendent nodes are at least* $\min(k_1, k_2)$*. Such a fact is utilized in the original LB algorithm.*

4. *Consider a node* (Ω_1, Ω_2) *and one of its children* (Ω_1', Ω_2')*. If* (Ω_1', Ω_2') *is the first child of* (Ω_1, Ω_2)*, then subset* Ω_1' *(resp.,* Ω_2'*) is a subset of* Ω_2 *by removing the last (resp. the second last) covariate from* Ω_2*. If* (Ω_1', Ω_2') *is not the first child, assuming* (Ω_3', Ω_4') *is the child of* (Ω_1, Ω_2) *and, in the pair tree, is immediately in the left hand side of* (Ω_1', Ω_2')*, then subset* Ω_1' *is obtained by removing the last covariate from subset* Ω_3' *and subset* Ω_2' *is obtained by removing one covariate from* Ω_2*. In summary, subsets of a particular node can be obtained by removing one covariate from a subset in its parent and possibly its left sibling. This relation ensures an efficient numerical approach to scan through all the nodes; more specifically, an efficient scan moves top down and left to right in the pair tree.*

7.3.3 Test in the Original Leaps-and-Bounds Algorithm

Now, after the analysis of the inverse tree and the pair tree, we are ready to give a new description of the LB method. We still use the case of 5 covariates as our example. Recall the contents of Section 7.2.2. The following inverse matrices can be computed, according to the scheme below:

$$\begin{aligned}
I(\{1\}) \rightarrow I(\{12\}) \rightarrow I(\{123\}) &\rightarrow I(\{1234\}) \rightarrow I(\{12345\}), \\
I(\{12345\}) &\rightarrow I(\{1235\}), \\
I(\{12345\}) &\rightarrow I(\{1245\}), \\
I(\{12345\}) &\rightarrow I(\{1345\}), \\
I(\{12345\}) &\rightarrow I(\{2345\}),
\end{aligned}$$

where each '→' involves inserting/removing one covariate to/from the subset on the left hand side. The consequent residual sums of squares can be computed correspondingly by

(65).

In the pair tree, we have the residual sums of squares of the subsets included in the root node and the nodes in the first level. We consider the remaining nodes. Whether or not to compute RSS($\{124\}$) and RSS($\{125\}$) depends on the values of RSS($\{1245\}$) and RSS($\{123\}$). If RSS($\{123\}$) $\leq$ RSS($\{1245\}$), because $\{124\}$ and $\{125\}$ are subsets of $\{1245\}$, we immediately have RSS($\{123\}$) $\leq$ RSS($\{124\}$) and RSS($\{123\}$) $\leq$ RSS($\{125\}$). Hence, there is no need to compute for RSS($\{124\}$) and RSS($\{125\}$). Otherwise, they should be computed.

Similarly, whether or not to compute RSS($\{134\}$) and RSS($\{135\}$) (or RSS($\{13\}$) and RSS($\{145\}$)) depends on three values: RSS($\{12\}$), RSS($\{1345\}$), and min(RSS($\{123\}$), RSS($\{124\}$), RSS($\{125\}$)) (denoted as $\underline{\text{RSS}}(3)$). Note that $\underline{\text{RSS}}(3) \leq$ RSS($\{12\}$). There are three cases for those three values:

- If RSS($\{12\}$) $\leq$ RSS($\{1345\}$), then none of RSS($\{134\}$), RSS($\{135\}$), RSS($\{13\}$), or RSS($\{145\}$) needs to be calculated.

- If $\underline{\text{RSS}}(3) \leq$ RSS($\{1345\}$) $<$ RSS($\{12\}$), then only RSS($\{13\}$) and RSS($\{145\}$) need to be calculated to update the minimum RSS with 2 covariates.

- If RSS($\{1345\}$) $< \underline{\text{RSS}}(3)$, then all of the four RSS's need to be calculated.

Repeating this step through the entire tree gives the original LB algorithm.

In general, the original LB algorithm is equivalent to scanning through the pair tree according to the following scheme.

- Compute the residual sums of squares for all the subsets in the root node and the nodes in level 1 of the pair tree.

- Suppose (Ω_1, Ω_2) is an intermediate node in the pair tree, and RSS(Ω_1) and RSS(Ω_2) have been computed. In our construction, readers can verify that we have $|\Omega_1| \leq |\Omega_2|$, where $|\cdot|$ is the size of a subset. For $|\Omega_1| \leq k \leq |\Omega_2|$, let $\underline{\text{RSS}}(k)$ denote the minimum of the residual sum of squares of all the k-subsets that have been scanned up to this point. If RSS(Ω_2) $\geq \underline{\text{RSS}}(k)$, for all $|\Omega_1| \leq k \leq |\Omega_2|$, then the computations for the descendants of node (Ω_1, Ω_2) can be ignored, because none of them can be an optimal solution to (FW). Otherwise, we should consider at least partial of the descendants of (Ω_1, Ω_2).

Readers can easily verify the following result.

Lemma 3.4 *In a top-down and left-to-right scheme to scan through the pair tree, the following inequality is true,*

$$\underline{RSS}(k+1) \leq \underline{RSS}(k),$$

for any k that is applicable.

Hence, in the LB algorithm, we have the following cases:

- If

  $$\underline{\mathrm{RSS}}(|\Omega_1|) \leq \mathrm{RSS}(\Omega_2),$$

  then skip all the descendants of node (Ω_1, Ω_2). Because none of the subsets in a descendant of node (Ω_1, Ω_2) can have a smaller residual sum of squares than the corresponding existing $\underline{\mathrm{RSS}}(k)$'s.

- If

  $$\underline{\mathrm{RSS}}(|\Omega_2| - k) \leq \mathrm{RSS}(\Omega_2) < \underline{\mathrm{RSS}}(|\Omega_2| - k - 1)$$

  for certain k, where $1 \leq k \leq |\Omega_2| - |\Omega_1| - 1$, then we can skip the first k children of node (Ω_1, Ω_2).

- If

  $$\mathrm{RSS}(\Omega_2) < \underline{\mathrm{RSS}}(|\Omega_2| - 1),$$

  then none of the children of (Ω_1, Ω_2) can be skipped.

In summary, the optimality tests in LB completely depends on the values of residual sums of squares.

7.4 Additional Optimality Tests

To the best of our knowledge, little effort has been reported to bring new optimality tests. In this section, additional tests are derived. The key intuition is to bring in the considerations of the coefficients and the residuals in the up-to-date optimal solutions. In comparison, the original LB method only considers the values of residual sums of squares. Additional optimality tests, together with the original test, will reduce the number of subsets that are needed to be considered. Hence, it reduces the computational requirement.

7.4.1 New Tests

We now consider additional optimality tests for node (Ω_1, Ω_2) in the pair tree. The following notations will be used:

- Let $\Omega_{(k)}$ be the k-subset associated with the minimum residual sum of squares $\underline{\mathrm{RSS}}(k)$.

- Let $\widehat{\beta}_{(k)} = \widehat{\beta}(\Omega_{(k)})$ denote the coefficients of the least square fit on the subset $\Omega_{(k)}$, whose computation is given in (64).

- Denote a residual vector
$$\varepsilon_{(k)} = y - \Phi_{\Omega_{(k)}} \widehat{\beta}_{(k)}.$$

- Recall (Ω_1, Ω_2) is a pair of subsets in the pair tree. Recall ϕ_a and ϕ_b are standardized covariates. A new quantity μ is defined as follows:
$$\mu = \max_{\substack{a,b \in \Omega_2 \cup \Omega_{(k)} \\ a \neq b}} |\langle \phi_a, \phi_b \rangle|.$$
Quantity μ is the maximum absolute value of the correlation within the subset $\Omega_2 \cup \Omega_{(k)}$.

- Define $k_1(k) = \min(2k, |\Omega_2 \cup \Omega_{(k)}|)$. Quantities μ and $k_1(k)$ are easily computable.

- For an arbitrary vector v and an arbitrary integer k_2, assuming that the dimension of v is no less than k_2, we define
$$\|v\|_{(k_2)} = \sqrt{\sum_{j=1}^{k_2} |v|_{(j)}^2},$$

where $|v|_{(1)} \geq |v|_{(2)} \geq |v|_{(3)} \geq \cdots$ are ordered absolute values of the entries of vector v.

- It is easy to observe that vector $\Phi^T \varepsilon_{(k)}$ is an $(m+1)$-dimensional vector, which is handly computable. We define vector $\Phi^T_{\Omega_2 \cup \Omega_{(k)}} \varepsilon_{(k)}$ as a subvector of $\Phi^T \varepsilon_{(k)}$ by taking covariate indices in the subset $\Omega_2 \cup \Omega_{(k)}$.

The following theorem points out a new optimality test.

Theorem 4.1 (Optimality Rule) *For the previously defined* $\Omega_1, \Omega_2, \Omega_{(k)}, \varepsilon_{(k)}, k_1(k)$ *(simplified as* k_1*) and* μ*. For* $|\Omega_1| \leq k \leq |\Omega_2|$*, define a set* $\Theta(k)$ *of covariate indices such that* $i \in \Theta(k)$ *if and only if* $i \in \Omega_{(k)}$ *and*

$$|(\widehat{\beta}_{(k)})_i| > \frac{\|\Phi^T_{\Omega_2 \cup \Omega_{(k)}} \varepsilon_{(k)}\|_\infty + \|\Phi^T_{\Omega_2 \cup \Omega_{(k)}} \varepsilon_{(k)}\|_{(k_1)}}{1 - (k_1 - 1)\mu}, \tag{66}$$

where $(\widehat{\beta}_{(k)})_i$ *denotes the coefficient of the ith covariate in* $\widehat{\beta}_{(k)}$*. If a subset* Ω *(*$\Omega \subset \Omega_2$ *and* $|\Omega| = k$*) achieves RSS*$(\Omega) \leq \underline{RSS}(k)$*, then we must have* $\Theta(k) \subset \Omega$*.*

If a k-subset achieves a residual sum of squares that is less than $\underline{\text{RSS}}(k)$, then $\Theta(k)$ is a subset of this k-subset. Hence, in the pair tree, any descendant that does not include $\Theta(k)$ as a subset cannot achieve a residual sum of squares less than $\underline{\text{RSS}}(k)$. This fact can be used to screen out some descendants of the nodes in a pair tree.

7.4.2 Proof of Theorem 4.1

The key idea adopted in the proof is to find a sufficient condition for a subset, such that this subset can not achieve a smaller residual sum of square than the one that corresponds to the up-to-date optimal subset having the same size.

We use the same notations as in the previous subsection. Let Ω be a subset of Ω_2: $\Omega \subset \Omega_2$. Let δ denote a vector that satisfies the following rules.

- δ only takes possibly nonzero values at position i when $i \in \Omega \cup \Omega_{(k)}$.

- Let $\widehat{\beta}(\Omega)$ denote the coefficient of the least square fit when the subset is Ω. Given previously defined $\widehat{\beta}_{(k)}$, the entries of δ are given as follows:

$$\delta_i = \begin{cases} (\widehat{\beta}(\Omega))_i - (\widehat{\beta}_{(k)})_i, & \text{if } i \in \Omega \cap \Omega_{(k)}, \\ (\widehat{\beta}(\Omega))_i, & \text{if } i \in \Omega, \text{ however } i \notin \Omega_{(k)}, \\ -(\widehat{\beta}_{(k)})_i, & \text{if } i \in \Omega_{(k)}, \text{ however } i \notin \Omega, \\ 0, & \text{elsewhere,} \end{cases}$$

where $(\cdot)_i$ denotes the value of the coefficient corresponding to the covariate i in the coefficient vector.

The following derives a necessary condition for RSS$(\Omega) \leq \underline{\text{RSS}}(k)$ =RSS$(\Omega_{(k)})$. We start with the following inequality:

$$\|y - \Phi_\Omega \widehat{\beta}(\Omega)\|_2^2 \leq \|\varepsilon_{(k)}\|_2^2.$$

The above is equivalent to the following:

$$\|\varepsilon_{(k)} - \Phi_{\Omega \cup \Omega_{(k)}} \delta\|_2^2 \leq \|\varepsilon_{(k)}\|_2^2,$$

which is equivalent to the following inequality:

$$\|\Phi_{\Omega \cup \Omega_{(k)}} \delta\|_2^2 \leq 2\langle \Phi_{\Omega \cup \Omega_{(k)}}^T \varepsilon_{(k)},\ \delta\rangle, \tag{67}$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product of two vectors.

In order to prove the theorem, we will need the following two lemmas.

Lemma 4.2 *Recall $|\Omega_{(k)}| = k$. Given $\Omega \subset \Omega_2$ and $|\Omega| = k$, we have*

$$|\Omega \cup \Omega_{(k)}| \leq \min(2k, |\Omega_2 \cup \Omega_{(k)}|) = k_1.$$

The proof of the above is simple, we leave it for the readers. The next result is critical in constructing the new optimality tests.

Lemma 4.3 *Recall $k_1 = \min(2k, |\Omega_2 \cup \Omega_{(k)}|)$. Given the previously defined quantities μ and δ, we have*

$$\|\Phi_{\Omega \cup \Omega_{(k)}} \delta\|_2^2 \geq (1 - (k_1 - 1)\mu)\, \|\delta\|_2^2.$$

Proof. First of all, we have

$$\begin{aligned}\|\Phi_{\Omega\cup\Omega_{(k)}}\delta\|_2^2 &= \delta^T\Phi_{\Omega\cup\Omega_{(k)}}^T\Phi_{\Omega\cup\Omega_{(k)}}\delta \\ &\geq \sum_i \delta_i^2 - \mu \sum_{a\neq b, a,b\in\Omega\cup\Omega_{(k)}} |\delta_a|\cdot|\delta_b|.\end{aligned}$$

Applying the Cauchy's inequality, readers can easily verify the following:

$$\sum_{\substack{a\neq b,\\ a,b\in\Omega\cup\Omega_{(k)}}} |\delta_a|\cdot|\delta_b| \leq (k_1-1)\|\delta\|_2^2.$$

Combining the above two, we have proved the inequality in the lemma. □

Combining Lemma 4.3 and inequality (67), we have

$$(1-(k_1-1)\mu)\,\|\delta\|_2^2 \leq 2\langle\Phi_{\Omega\cup\Omega_{(k)}}^T\varepsilon_{(k)},\ \delta\rangle,$$

which is equivalent to

$$\left\|(1-(k_1-1)\mu)\,\delta - \Phi_{\Omega\cup\Omega_{(k)}}^T\varepsilon_{(k)}\right\|_2^2 \leq \|\Phi_{\Omega\cup\Omega_{(k)}}^T\varepsilon_{(k)}\|_2^2.$$

The above implies the following,

$$(1-(k_1-1)\mu)\,\|\delta\|_\infty \leq \|\Phi_{\Omega\cup\Omega_{(k)}}^T\varepsilon_{(k)}\|_\infty + \|\Phi_{\Omega\cup\Omega_{(k)}}^T\varepsilon_{(k)}\|_2, \tag{68}$$

where $\|\cdot\|_\infty$ denotes the maximum absolute value in a vector. Recalling $\Omega\subset\Omega_2$, the following is evident:

$$\|\Phi_{\Omega\cup\Omega_{(k)}}^T\varepsilon_{(k)}\|_\infty \leq \|\Phi_{\Omega_2\cup\Omega_{(k)}}^T\varepsilon_{(k)}\|_\infty. \tag{69}$$

It is easy to verify that

$$\|\Phi_{\Omega\cup\Omega_{(k)}}^T\varepsilon_{(k)}\|_2 \leq \|\Phi_{\Omega_2\cup\Omega_{(k)}}^T\varepsilon_{(k)}\|_{(k_1)}. \tag{70}$$

Combining (69) and (70), we have

$$(1-(k_1-1)\mu)\,\|\delta\|_\infty \leq \|\Phi_{\Omega_2\cup\Omega_{(k)}}^T\varepsilon_{(k)}\|_\infty + \|\Phi_{\Omega_2\cup\Omega_{(k)}}^T\varepsilon_{(k)}\|_{(k_1)}. \tag{71}$$

Given (71), we are ready to prove the theorem. For $i\in\Omega_{(k)}$, suppose (66) holds. From (71), we have

$$|(\widehat{\beta}_{(k)})_i - (\widehat{\beta}(\Omega))_i| \leq \frac{\|\Phi_{\Omega_2\cup\Omega_{(k)}}^T\varepsilon_{(k)}\|_\infty + \|\Phi_{\Omega_2\cup\Omega_{(k)}}^T\varepsilon_{(k)}\|_{(k_1)}}{1-(k_1-1)\mu}.$$

The above and (66) lead to $|(\widehat{\beta}(\Omega))_i| > 0$, which implies that $i\in\Omega$. Hence, we have $\Theta(k)\subset\Omega$. The theorem is proven.

7.5 *Algorithm*

In this section, the implementation strategies are described. The scanning described in Section 7.5.1 is equivalent to the method in the original LB [28]. We believe our new description is more accessible. The integration of new optimality tests is trivial. Hence, it is only briefly described in Section 7.5.2.

7.5.1 A Scheme to Scan Through the Pair Tree $PT(m)$

We design an algorithm that reaches each node of $PT(m)$ once and only once. We will use the following notations. For an arbitrary set Ω with ordered elements, for integer $j \geq 1$, $r(\Omega, j)$ denotes a subset of Ω by removing the jth last element of Ω. For example, we have

$$r(\{12345\}, 1) = \{1234\}, \text{ and } r(\{12345\}, 5) = \{2345\}.$$

Based on the above, define

$$r^k(\Omega, j) = \underbrace{r(r(\cdots r}_{k \text{ times}} \ (\Omega, j) \cdots, j), j).$$

For example, we have $r^2(\{12345\}, 1) = \{123\}$ and $r^3(\{12345\}, 1) = \{12\}$.

Given the structure of $PT(m)$, readers can verify that the following scheme reaches every node in $PT(m)$ once and only once.

1. A *node list* is empty initially. Starting from the root node $(\emptyset, \{1, 2, \ldots, m\})$, the following array,

$$\begin{array}{ccc}
r(\{1,2,\ldots,m\},1), & r(\{1,2,\ldots,m\},2), & 1, \\
r^2(\{1,2,\ldots,m\},1), & r(\{1,2,\ldots,m\},3), & 2, \\
r^3(\{1,2,\ldots,m\},1), & r(\{1,2,\ldots,m\},4), & 3, \\
\vdots & \vdots & \vdots \\
r^{m-1}(\{1,2,\ldots,m\},1), & r(\{1,2,\ldots,m\},m), & m-1,
\end{array}$$

   is inserted into the node list. Note each row of the node list is made by two subsets and its order among the siblings.

2. Suppose the node list is not empty and the top row is (Ω_1, Ω_2, s), where Ω_1 and Ω_2 are the subsets of $\{1, \ldots, m\}$, and integer $s \geq 1$.

   - If $s = 1$, this row is removed from the node list (has been scanned).
   - If $s > 1$, add the following array at the bottom of the node list:

$$\begin{array}{ccc} r(\Omega_2, 1), & r(\Omega_2, 2), & 1, \\ r^2(\Omega_2, 1), & r(\Omega_2, 3), & 2, \\ \vdots & \vdots & \vdots \\ r^{s-1}(\Omega_2, 1), & r(\Omega_2, s), & s-1. \end{array}$$

   Then, remove the top row (Ω_1, Ω_2, s) from the node list.

3. Step 2 is repeated until the node list is empty again.

7.5.2 Integrating the Optimality Tests

In the scheme that was described in the last subsection, it is straightforward to integrate the optimality test. We maintain the *optimality tests list*, whose rows are made by

$$\underline{\mathrm{RSS}}(k),\ \Omega_{(k)},\ \widehat{\beta}_{(k)}^T,\ \varepsilon_{(k)}^T \Phi,$$

where $0 \leq k \leq m$. In step 2 in the last subsection, if the pair of subsets in a row fail at least one optimality test (which includes both the original test in LB and the newly proposed test in Theorem 4.1), then this row is not inserted into the node list.

Note the original LB only uses the information in the first column of the optimality tests list, while the newly proposed tests (LBOT) use additional information.

7.6 *Simulations*

7.6.1 Synthetic Data

7.6.1.1 An Illustrative Example

In order to illustrate the efficiency of our method, we create a table – Table 7.6.1.1 – that includes all the pairs from $PT(5)$. The pairs that are used by the original LB are marked with '$*$', while those that are required by our enhanced leaps-and-bounds algorithm are marked with 'Δ'. The underlying regression model is:

$$y = 167.5058 + 27.0171x_1 + 5.2054x_3 + 135.8065x_4 - 0.0431x_5 + \varepsilon,$$

which is generated by random. It is observed that for the case illustrated in Table 7.6.1.1, the enhanced leaps-and-bounds reduces the number of examined pairs nodes from 13 (which is for the original LB) to 9. Table 7.6.1.1 presents the minimum residual sums of squares for different subset size k, as well as the optimal k-subsets.

Table 2: Pair tree and residual sums of squares.

LBOT	LB	Ω_1	RSS	Ω_2	RSS
Δ	$*$	$\emptyset$	20030.67	12345	1077.78
Δ	$*$	1234	1077.81	1235	19667.34
Δ	$*$	123	19667.86	1245	1104.57
Δ	$*$	12	19691.99	1345	1077.80
Δ	$*$	1	19733.12	2345	1731.38
	$*$	124	1104.59	125	19691.40
	$*$	134	1077.83	135	19708.79
Δ	$*$	13	19709.37	145	1104.61
		234	1731.45	235	19964.93
Δ	$*$	23	19965.92	245	1759.68
Δ	$*$	2	19991.05	345	1731.45
	$*$	14	1104.62	15	19732.47
		24	1759.77	25	19989.98
		34	1731.51	35	20004.86
Δ	$*$	3	20005.92	45	1759.76
	$*$	4	1759.85	5	20029.52

Table 3: Minimum residual sums of squares and the corresponding optimal k-subsets.

k	RSS	Optimal Subset
0	20030.67	$\emptyset$
1	1759.85	4
2	1104.62	14
3	1077.83	134
4	1077.80	1345
5	1077.78	12345

7.6.1.2 LB versus LBOT in Random Experiments

To further compare LBOT with LB, some random experiments are performed. Recall the regression model, $y = \Phi x + \varepsilon$, where model matrix $\Phi \in \mathbb{R}^{n\times(m+1)}$. In the following experiments, we set $n = 1000$ and $m = 10$. Each column $\phi_i, i = 1, 2, \ldots, m$, is first generated from multivariate normal $N(\overrightarrow{0}_{n\times1}, I_n)$, then followed by standardization. Coefficients are generated as $x_i \sim N(0, \sigma^2), i = 0, 1, \ldots, m$, where $\sigma = 100$. Set $\varepsilon_i \sim N(0, 1)$, for $i = 1, 2, \ldots, n$. We present the dot-plots of quantities –

(number of pairs used in LBOT, number of pairs used in LB)

– in Figure 58. Totally 1000 random simulations are performed. The dashed line is the diagonal. We can see that all points are below the diagonal: LBOT always requires less pairs to be examined. To illustrate the reduction of the number of pairs of subsets that are required to be examined, in Figure 59, we present the histograms of the ratios – the number of pairs used in LBOT over the number of pairs used in leaps and bounds. On average, LBOT requires 87.05% of the subsets that are required by LB.

In Table 7.6.1.2, for a range of the values of m, following the same random simulations described in the last paragraph, the average number of pairs that are examined based on 10 random experiments are reported. Again, we see a reduction in the number of required pairs. It is interesting to observe that the percentage of pairs that are examined in a pair tree reduces as the number of covariates (m) increases; see the last column of Table 7.6.1.2.

Table 4: For different values of m, the average numbers of pairs that are examined by both LB and LBOT among ten random experiments are presented. The second column includes the total number of pairs in the complete pair trees.

m	# Pairs	LB	LBOT	LBOT/#Pairs
10	512	148.7	131.3	0.2564
12	2048	598.9	550.6	0.2688
14	8192	1714.3	1643.3	0.2006
16	32768	6226.2	6112.2	0.1865
18	262144	21034.0	20973.7	0.0800
20	1048576	42654.8	42278.6	0.0403

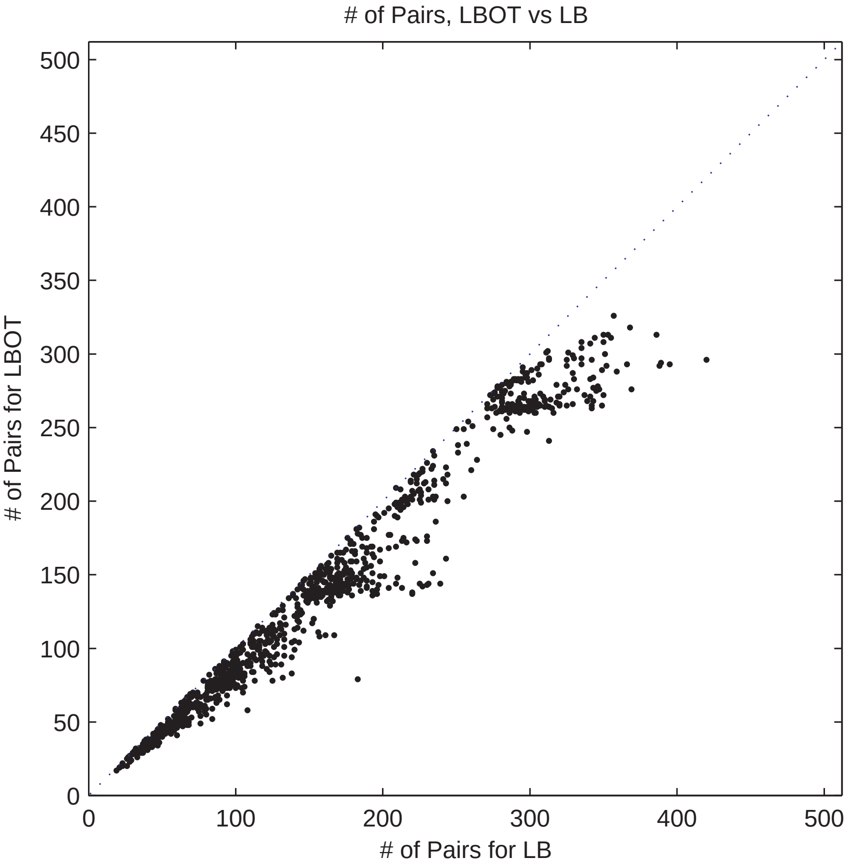


Figure 58: The number of pairs in LBOT versus the number of pairs used in LB. The number of covariates $m = 10$.

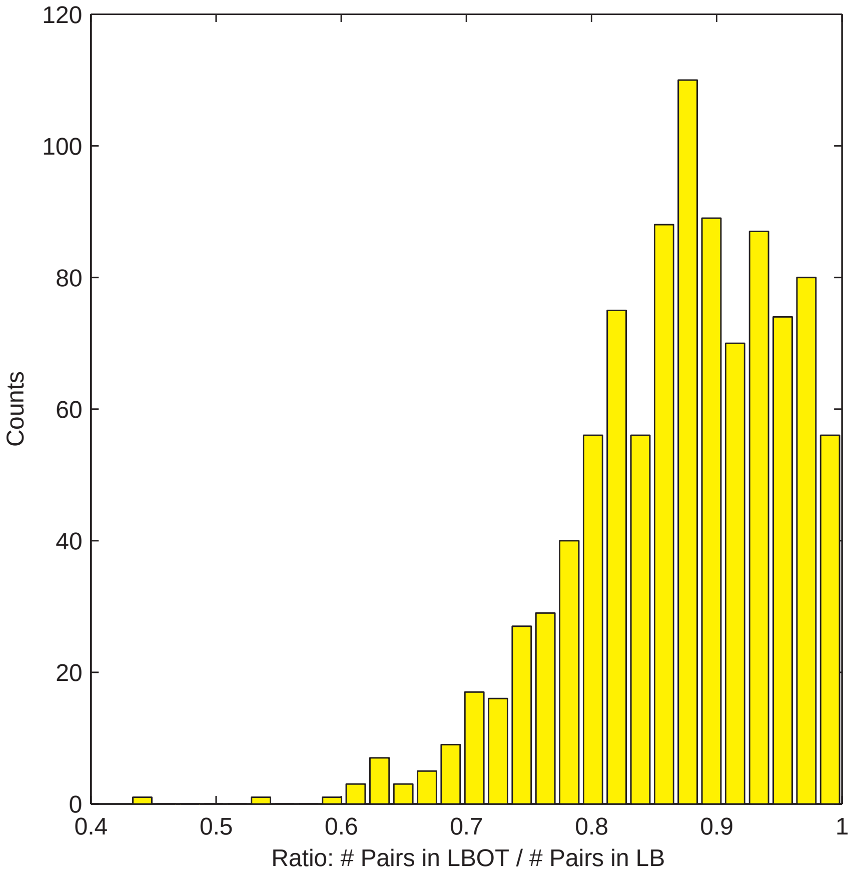


Figure 59: Histogram: when $m = 10$, the number of pairs in LBOT over the number of pairs used in LB.

7.6.2 Effect of Model

For simplicity, the experiment that leads to Figures 58 and 59 is denoted as Exp. A. Note in this experiment, the coefficients are sampled from $N(0, 100^2)$; i.e., the absolute values of the coefficients are likely to be large. To see when LBOT can significantly reduce the number of pairs from LB, we repeat the Exp. A, but change the distribution from $N(0, 100^2)$ to $N(0, 1)$. The new experiment is denoted as Exp. B. Figure 60 presents the histograms of the numbers of pairs used by LB, together with the ratios of the pairs between the two methods. Note Figure 60 (b) repeats Figure 59. It is redrawn and scaled for comparison. No dramatic difference can be seem from Figure 60 (a) and (c): the number of pairs that are required by LB does not change much. However, being compared with Figure 60 (b), histogram (d) is significantly skewed to the right: the additional optimality tests are less likely to reduce the computation in Exp. B, in which the coefficients of the underlying models are likely to be close to zero.

In summary, the additional optimality tests are likely to improve LB when the underlying true model has relatively large absolute values of coefficients (relative to the noise level).

7.6.3 Heuristic: Pre-Sorting

The pair tree is scanned top-down, left-to-right. If the optimal subsets appear earlier in the scanning scheme, LB and LBOT will have a better chance to reduce the computation. Based on this observation, we can reorder the covariates before the evocation of LB or LBOT. We carry out a random experiment (denoted as Exp. C), which is identical with Exp. A, except a pre-sorting of the covariates that imposes the following condition:

$$\langle y, \phi_1 \rangle \geq \langle y, \phi_2 \rangle \geq \cdots \geq \langle y, \phi_m \rangle.$$

The histograms of the number of pairs that are used by LB in Exp. A and C are presented in Figure 61 (a) and (c), respectively. After the pre-sorting, we see a significant reduction in the numbers of pairs that are required. Figure 61 (b) is another rescaled version of Figure 59. It will be compared with subfigure (d). Based on Figure 61 (d), LBOT still reduces the number of pairs in a significant proportion of cases.

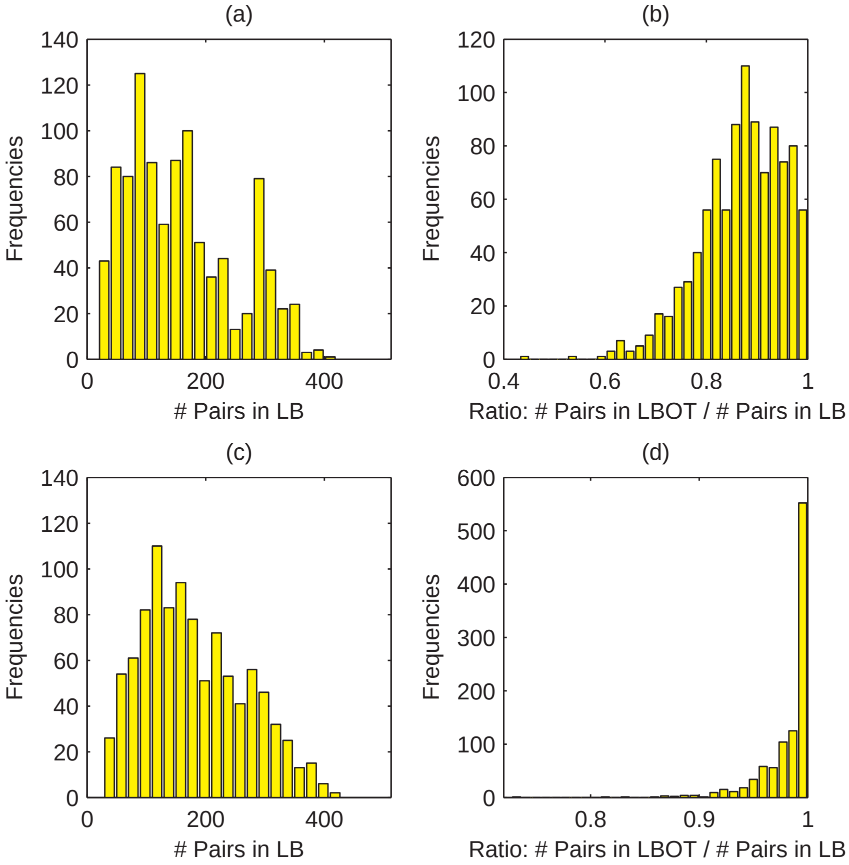


Figure 60: Histograms: (a) and (c), the number of pairs in LB; (b) and (d), ratios between the numbers of pairs in LB and the numbers of pairs in LBOT. (a) and (b) (resp., (c) and (d)) are for Exp. A (resp., Exp. B). See the context for details.

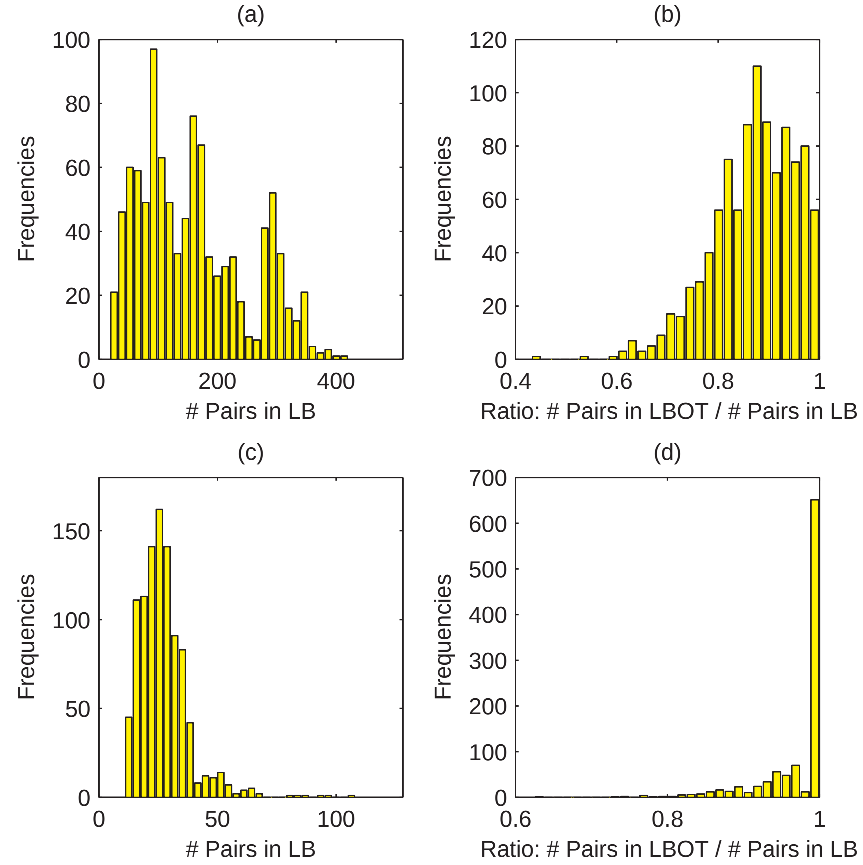


Figure 61: Histograms: (a) and (c), the number of pairs in LB; (b) and (d), ratios between the numbers of pairs in LB and the numbers of pairs in LBOT. Subfigures (a) and (b) (resp., (c) and (d)) are for Exp. A (resp., Exp. C). See the context for details.

In summary, some preprocessing can help to improve the efficiency of both LB and LBOT.

7.6.4 Real Data

Being compared with LB, LBOT always reduces the number of subsets that are required to be examined. The actual amount of reduction depends on the data, as one can observe in the previous experiments. For example, the reduction is more evident in Exp. A than in Exp. B. How much will LBOT reduce the number of subsets in real data? We experimented with two datasets: diabetes data (http://www-stat.stanford.edu/∼hastie/Papers/LARS/) and housing data (http://www.ics.uci.edu/∼mlearn/databases/housing/). They are chosen because they have been widely used in the regression literature. The diabetes data has 10 covariates and has been used in [25] to illustrate a stepwise algorithm (LARS). The housing data has 13 covariates. In both cases, LBOT fails to reduce the number of subsets. In other words, it is equivalent to the case when the ratio is equal to 1 in Exp. A, B, or C. Note in the random experiments, comparing to the histograms in Figure 59, 60 (d), and 61 (d), the probability of having a ratio "1" is small. To our surprise, the ratio is "1" for both 'real' data sets that we experimented. As a future research topic, it will be interesting to derive reasonable condition(s), under which the additional optimality tests can *not* reduce the number of required pairs in LB (or LBOT).

For the diabetes data, Table 7.6.4 gives the optimal subsets. The covariates are standardized. No pre-sorting is adopted. Note that some covariates come and leave the optimal subsets, e.g., covariates 5 and 7, as the subset size increases. Such a phenomenon can not be caught by a pure forward or backward subset selection.

For the housing data, the same table is produced in Table 7.6.4. Again, we see some covariates (e.g., 2 and 4) coming and leaving from the optimal subsets.

7.7 Discussion

LB is a branch-and-bound (B&B) approach designed specifically for regression problems. B&B has been applied to many other problems, e.g., *feature subset selection* (FSS). Typical references are [70], [58], and [54]. The objective in an FSS problem is different from the

Table 5: The optimal subsets and corresponding residual sums of squares, for the diabetes data.

Subset Sizes k	RSS	Optimal Subsets									
0	2621009.12										
1	1719581.81			3							
2	1416694.01			3						9	
3	1362708.69			3	4					9	
4	1331431.40			3	4	5				9	
5	1287881.16		2	3	4			7		9	
6	1271494.00		2	3	4	5	6			9	
7	1267807.81		2	3	4	5	6		8	9	
8	1264714.58		2	3	4	5	6		8	9	10
9	1264068.10		2	3	4	5	6	7	8	9	10
10	1263985.79	1	2	3	4	5	6	7	8	9	10

Table 6: The optimal subsets and corresponding residual sums of squares for the Housing data.

Subset Sizes k	RSS	Optimal Subsets												
0	42716.30													
1	19472.38													13
2	15439.31						6							13
3	13727.99						6					11		13
4	13228.91						6		8			11		13
5	12469.34					5	6		8			11		13
6	12141.07				4	5	6		8			11		13
7	11868.24				4	5	6		8			11	12	13
8	11678.30		2		4	5	6		8			11	12	13
9	11526.12	1			4	5	6		8	9		11	12	13
10	11308.58	1	2			5	6		8	9	10	11	12	13
11	11081.36	1	2		4	5	6		8	9	10	11	12	13
12	11078.85	1	2	3	4	5	6		8	9	10	11	12	13
13	11078.78	1	2	3	4	5	6	7	8	9	10	11	12	13

objective in a regression problem. FSS mostly involves with a classification problem, instead of a regression problem. In FSS, researchers have proposed enhanced B&B by adapting various heuristics: [99], [13], [88], and [9]. Similarly, leaps-and-bounds has been applied to variable selection in discriminant analysis [87]. A careful comparison will review that the strategy of deriving the additional optimality test in LBOT is very distinct from the above works in FSS. On the other hand, it will be very interesting to explore the heuristics that have been developed in accelerating the B&B algorithms in FSS. They could further improve the performance of LBOT. Some serious structural works are required.

7.8 Conclusion

In this chapter, new optimality conditions are derived in the framework of leaps-and-bounds algorithm. The new tests guarantee to reduce the number of subsets that are required in a branch-and-bound exhaustive subset search. The reduction of computation is testified in random experiments. We improved a state-of-the-art method in comprehensive subset selections. The ideas behind the newly introduced tests are novel. The analysis technique that is used in deriving the new condition could be insightful in studying other regression-related problems.

APPENDIX A

PROOFS ASSOCIATED WITH CHAPTER VI (PART II)

A.1 Details for Proving (26)

We have

$$
\begin{aligned}
I(f_0) &= I(f_\theta)|_{\theta=0} \\
&= \int \left(\frac{\partial}{\partial\theta} f_\theta\right)^2 \frac{1}{f_\theta} dx|_{\theta=0} \\
&\overset{(16)}{=} \int_{-\delta}^{\delta} \left(2\cdot c\cdot \cos\lambda_1 \frac{x}{\delta} \cdot \frac{\lambda_1}{\delta} \cdot \sin\frac{\lambda_1 x}{\delta}\right)^2 \frac{1}{c\left[\cos\lambda_1 \frac{x}{\delta}\right]^2} dx \\
&\quad +2\int_{\delta}^{+\infty} \frac{\left[c\cdot \exp\left(-2\lambda_2 \frac{x}{\delta}\right)\cdot \frac{2\lambda_2}{\delta}\cdot \cos^2\lambda_1 \cdot \exp(2\lambda_2)\right]^2}{c\cdot \exp\left(-2\lambda_2\frac{x}{\delta}\right)\cdot\cos^2\lambda_1\cdot\exp(2\lambda_2)} dx \\
&= 4c\frac{\lambda_1^2}{\delta^2}\int_{-\delta}^{\delta} \sin^2\frac{\lambda_1 x}{\delta} dx \\
&\quad +2c\frac{4\lambda_2^2}{\delta^2}\cos^2\lambda_1\cdot\exp(2\lambda_2)\int_{\delta}^{+\infty}\exp\left(-2\lambda_2\frac{x}{\delta}\right)dx \\
&= 4c\frac{\lambda_1^2}{\delta}\left(1-\frac{1}{2\lambda_1}\sin 2\lambda_1\right) \\
&\quad +4c\frac{\lambda_2}{\delta}\cos^2\lambda_1 \\
&\overset{(21)}{=} 4c\frac{\lambda_1^2}{\delta} \overset{(23)}{=} 4\frac{\lambda_1^2}{\delta^2}\frac{\alpha\lambda_2}{\cos^2\lambda_1} \overset{(24)}{=} 4\frac{\lambda_1^2}{\delta^2}\frac{\lambda_2\cos\lambda_1}{\lambda_1\cdot\sin\lambda_1+\cos\lambda_1} \\
&\overset{(21)}{=} 4\frac{\lambda_1^2}{\delta^2}\frac{\lambda_1\cdot\sin\lambda_1}{\lambda_1\cdot\sin\lambda_1+\cos\lambda_1}.
\end{aligned}
$$

A.2 Some Lemmas Regarding Cut-off Values in Section 5.6.2.2

Lemma 2.1 *Random variable $\|d_{LS,1}\|_2^2$ satisfies the χ_m^2 distribution with m degrees of freedom, where m is the column rank of matrix A.*

Proof. Based on (35),

$$d_{LS,1} = \mathbf{P}_{A,LS}(\varepsilon) = A(A^TA)^{-1}A^T\varepsilon.$$

Since $A(A^TA)^{-1}A^T$ is a projection matrix with the rank equals to m, it can be written as

$$A(A^TA)^{-1}A^T = O \cdot \begin{pmatrix} \mathbf{I}_m & 0 \\ 0 & 0 \end{pmatrix}_{n\times n} \cdot O^T,$$

where O is an orthogonal matrix. Hence

$$\|d_{LS,1}\|_2^2 = \left\| \begin{pmatrix} \mathbf{I}_m & 0 \\ 0 & 0 \end{pmatrix} \cdot O^T \cdot \varepsilon \right\|_2^2.$$

Since $O^T \cdot \varepsilon$ also satisfies standard Normal distribution in $\mathbf{R}^n$, i.e., $O^T \cdot \varepsilon \sim \text{Normal}(0, \mathbf{I}_n)$, we have

$$\|d_{LS,1}\|_2^2 \sim \chi_m^2.$$

□

Lemma 2.2 *The ratio $\|d_{LS,1}\|_\infty / \|d_{LS,1}\|_2$ has the same distribution as random variable $\rho_{max,m}$, which was defined in Section 5.6.2.2.*

Proof. Let $\eta = d_{LS,1}/\|d_{LS,1}\|_2$. Recall

$$d_{LS,1} = \mathbf{P}_{A,LS}(\varepsilon) = U^T \begin{pmatrix} \mathbf{I}_m & \\ & 0 \end{pmatrix} U \cdot \varepsilon.$$

Because $\varepsilon \sim N(0, \mathbf{I}_n)$, we have $U \cdot \varepsilon \sim N(0, \mathbf{I}_n)$ as well. Hence

$$\begin{aligned} \frac{d_{LS,1}}{\|d_{LS,1}\|_2} &= \frac{U^T \begin{pmatrix} \mathbf{I}_m & \\ & 0 \end{pmatrix}_{n\times n} U \cdot \varepsilon}{\| \begin{pmatrix} \mathbf{I}_m & \\ & 0 \end{pmatrix}_{n\times n} U \cdot \varepsilon\|_2} \\ &= U^T \begin{pmatrix} \tilde{x}_m / \|\tilde{x}_m\|_2 \\ 0_{(n-m)\times 1} \end{pmatrix}, \end{aligned}$$

where $\tilde{x}_m = (\mathbf{I}_m \quad 0)_{m\times n} U \cdot \varepsilon \sim N(0, \mathbf{I}_m)$. Based on the property of a normally distributed random vector, we have $x_m = \tilde{x}_m / \|\tilde{x}_m\|_2$ is uniformly distributed on the unit sphere $S^{m-1} \subset \mathbf{R}^m$. □

Lemma 2.3 *Random variables* $\|d_{LS,1}\|_2$ *and* $\|d_{LS,1}\|_\infty/\|d_{LS,1}\|_2$ *are independent.*

Proof. Let $\eta = d_{LS,1}/\|d_{LS,1}\|_2$. From the previous lemma, the distribution of the random variable η is independent of the quantity $\|d_{LS,1}\|_2$. Hence $\|d_{LS,1}\|_\infty/\|d_{LS,1}\|_2 = \|\eta\|_\infty$ is independent of $\|d_{LS,1}\|_2$. □

APPENDIX B

PROOFS ASSOCIATED WITH CHAPTER VII (PART II)

B.1 Proof of Lemma 2.2

Suppose there are two vectors x_1 and x_2, such that $P_1 = (\|x_1\|_1, \|y - \Phi x_1\|_2^2)$, $P_2 = (\|x_2\|_1, \|y - \Phi x_2\|_2^2)$, and for $0 < \lambda < 1$, $\lambda P_1 + (1-\lambda)P_2$ are on the frontier. We consider a point $P_\lambda = (\lambda x_1 + (1-\lambda)x_2\|_1, \|y - \Phi[\lambda x_1 + (1-\lambda)x_2]\|_2^2)$. First of all, we have inequality

$$\|\lambda x_1 + (1-\lambda)x_2\|_1 \leq \lambda\|x_1\|_1 + (1-\lambda)\|x_2\|_1, \tag{72}$$

and the equality holds if and only if x_1 and x_2 have the same sign at each position, or one of them takes zero. On the other hand, we have inequality

$$\|\lambda(y - \Phi x_1) + (1-\lambda)(y - \Phi x_2)\|_2^2 \leq \lambda\|y - \Phi x_1\|_2^2 + (1-\lambda)\|y - \Phi x_2\|_2^2, \tag{73}$$

and equality holds if and only if $y - \Phi x_1 = y - \Phi x_2$, which implies

$$\|y - \Phi x_1\|_2^2 = \|y - \Phi x_2\|_2^2. \tag{74}$$

It is impossible to have (74). The reason is the following. Given formula (72) and (73), and frontier being a non-increasing function on c, we can easily verify that the frontier is convex. If (74) is true, the three points P_1, P_2, and $(\|\tilde{x}\|_1, 0)$ will violate the convexity. Hence, $\lambda P_1 + (1-\lambda)P_2$ cannot be on the frontier. From all the above, we proved that the frontier is *strictly* convex.

B.2 Proof of Lemma 3.1

The choice of c_1 depends on the following three correlations:

1. For $m - A + 1 \leq i \leq m$,

$$\begin{aligned} \langle \phi_i, s - c_1\phi_1 \rangle &= \langle \delta_i, s - c_1(a_1 s + b_1\delta_1) \rangle \\ &= (1 - c_1 a_1)/\sqrt{A}. \end{aligned} \tag{75}$$

2. For $1 \leq j \leq m - A$,

$$\begin{aligned} \langle \phi_j, s - c_1\phi_1 \rangle &= \langle a_j s + b_j \delta_j, s - c_1(a_1 s + b_1 \delta_1) \rangle \\ &= a_j(1 - c_1 a_1) - c_1 b_j b_1 \langle \delta_j, \delta_1 \rangle. \end{aligned}$$

As special cases, one has

a. for $j = 1$,

$$\langle \phi_1, s - c_1\phi_1 \rangle = a_1 - c_1; \tag{76}$$

b. For $j \geq 2$,

$$\langle \phi_j, s - c_1\phi_1 \rangle = a_j(1 - c_1 a_1). \tag{77}$$

The choice of c_1 is the maximum value that satisfies (76) $\geq$ (75) and (76) $\geq$ (77). From (76) $\geq$ (75), we have $a_1 - c_1 \geq (1 - c_1 a_1)/\sqrt{A}$, which is equivalent to

$$a_1 - 1/\sqrt{A} \geq c_1(1 - a_1/\sqrt{A}). \tag{78}$$

From (76) $\geq$ (77), we have $a_1 - c_1 \geq a_j(1 - c_1 a_1)$, which is equivalent to

$$a_1 - a_j \geq c_1(1 - a_1 a_j). \tag{79}$$

Combining (78) and (79), we have

$$\begin{aligned} c_1 &= \min \left\{ \frac{a_1 - 1/\sqrt{A}}{1 - a_1/\sqrt{A}}, \frac{a_1 - a_j}{1 - a_1 a_j} \right\} \\ &= \frac{a_1 - a_2}{1 - a_1 a_2}. \end{aligned}$$

The last equality is based on an observation that function $\frac{a_1 - x}{1 - x a_1}$ is a decreasing function of x. □

B.3 Proof of Theorem 3.2

To prove the above theorem, we will need the following lemma.

Lemma 3.1 *The equiangular vector among $\phi_1, \phi_2, ..., \phi_k$ is*

$$u_k = [\phi_1 \ \phi_2 \ \cdots \ \phi_k] \left(D_k^{-1}\mathbf{1} - D_k^{-1} v_k \frac{v_k^T D_k^{-1}\mathbf{1}}{1 + v_k^T D_k^{-1} v_k} \right),$$

where

$$D_k = \begin{pmatrix} b_1^2 & 0 & \cdots & 0 \\ 0 & b_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & b_k^2 \end{pmatrix}, \qquad v_k = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_k \end{pmatrix},$$

and **1** *is a k-dimensional all one vector.*

Proof: It is easy to verify that

$$[\phi_1 \ \phi_2 \ \cdots \ \phi_k]^T [\phi_1 \ \phi_2 \ \cdots \ \phi_k] = D_k + v_k v_k^T.$$

Using a known result in linear algebra:

$$(D_k + v_k v_k^T)^{-1} = D_k^{-1} - D_k^{-1} v_k v_k^T D_k^{-1} \frac{1}{1 + v_k^T D_k^{-1} v_k}.$$

Denoting $\Phi_k = [\phi_1 \ \phi_2 \ \cdots \ \phi_k]$, we have

$$\begin{aligned} u_k &= \Phi_k (\Phi_k^T \Phi_k)^{-1} \mathbf{1} \\ &= \Phi_k \left(D_k^{-1} \mathbf{1} - D_k^{-1} v_k v_k^T D_k^{-1} \mathbf{1} \frac{1}{1 + v_k^T D_k^{-1} v_k} \right). \end{aligned}$$

□

Note that in order to keep the formula simple, we do not normalize the vector u_k. In LARS, this does not change the selection of variables.

Proof of Theorem 3.2 Now we apply induction to prove the theorem. Assume after step $k-1$, $k \leq m-A$, the variates $\phi_1, \phi_2, \ldots, \phi_k$ have been selected, and a surrogate residual is

$$\widetilde{r}_{k-1} = s - \sum_{j=1}^{k-1} \frac{a_j - a_k}{b_j} \delta_j.$$

We will argue that in the next step, variate ϕ_{k+1} will be chosen, and the next surrogate residual has the form

$$\widetilde{r}_k = s - \sum_{j=1}^{k} \frac{a_j - a_{k+1}}{b_j} \delta_j. \tag{80}$$

Combining the above two, the theorem is proven. We first perform a sanity check:

1. For $m - A + 1 \leq i \leq m$, $\langle \phi_i, \widetilde{r}_{k-1} \rangle = 1/\sqrt{A}$.

2. For $k \le j \le k$,

$$\begin{aligned}
\langle \phi_j, \widetilde{r}_{k-1} \rangle &= \langle a_j s + b_j \delta_j, s - \sum_{t=1}^{k-1} \frac{a_t - a_k}{b_t} \delta_t \rangle \\
&= a_j - (a_j - a_k) \\
&= a_k.
\end{aligned}$$

3. For $k + 1 \le j \le m - A$,

$$\langle \phi_j, \widetilde{r}_{k-1} \rangle = a_j.$$

The next residual should be

$$r_k = \widetilde{r}_{k-1} - c_k u_k,$$

where c_k is determined by considering the following three correlations:

1. For $i \ge m - A + 1$,

$$\begin{aligned}
\langle \phi_i, r_k \rangle &= \langle \phi_i, \widetilde{r}_{k-1} - c_k u_k \rangle \\
&= \langle \phi_i, s - \sum_{t=1}^{k-1} \frac{a_t - a_k}{b_t} \delta_t - c_k u_k \rangle \\
&= \frac{1}{\sqrt{A}} - c_k \langle \phi, u_k \rangle \\
&= \frac{1}{\sqrt{A}} - c_k \frac{1}{\sqrt{A}} v_k^T \left(D_k^{-1} \mathbf{1} - D_k^{-1} v_k v_k^T D_k^{-1} \mathbf{1} \frac{1}{1 + v_k^T D_k^{-1} v_k} \right) \\
&= \frac{1}{\sqrt{A}} \left[1 - c_k \frac{v_k^T D_k^{-1} \mathbf{1}}{1 + v_k^T D_k^{-1} v_k} \right] \\
&= \frac{1}{\sqrt{A}} [1 - c_k g(k)], \qquad (81)
\end{aligned}$$

where $g(k) = \frac{v_k^T D_k^{-1} \mathbf{1}}{1 + v_k^T D_k^{-1} v_k}$, and v_k, D_k, and $\mathbf{1}$ are defined in Lemma 3.1. This quantity will appear frequently in the following.

2. For $j \le k$,

$$\begin{aligned}
\langle \phi_j, r_k \rangle &= \langle a_j s + b_j \delta_j, s - \sum_{t=1}^{k-1} \frac{a_t - a_k}{b_t} \delta_t - c_k u_k \rangle \\
&= a_k - c_k \langle a_j s + b_j \delta_j, u_k \rangle.
\end{aligned}$$

From the definition of u_k, one has

$$\langle a_j s + b_j \delta_j, u_k \rangle = \langle \phi_j, u_k \rangle = 1.$$

Hence,

$$\langle \phi_j, r_k \rangle = a_k - c_k. \tag{82}$$

3. For $k+1 \le j \le m - A$,

$$\begin{aligned}
\langle \phi_j, r_k \rangle &= \langle a_j s + b_j \delta_j, s - \sum_{t=1}^{k-1} \frac{a_t - a_k}{b_t} \delta_t - c_k u_k \rangle \\
&= a_j - c_k \langle a_j s + b_j \delta_j, u_k \rangle \\
&= a_j - c_k a_j \langle s, u_k \rangle \\
&= a_j - c_k a_j v_k^T \left(D_k^{-1} \mathbf{1} - D_k^{-1} v_k v_k^T D_k^{-1} \mathbf{1} \frac{1}{1 + v_k^T D_k^{-1} v_k} \right) \\
&= a_j - c_k a_j \frac{v_k^T D_k^{-1} \mathbf{1}}{1 + v_k^T D_k^{-1} v_k} \\
&= a_j [1 - c_k g(k)]. \qquad (83)
\end{aligned}$$

To determine c_k, we consider two conditions: (82) $\ge$ (81) and (82) $\ge$ (83). From (82) $\ge$ (81), we have $a_k - c_k \ge \frac{1}{\sqrt{A}}[1 - c_k g(k)]$, which is equivalent to

$$a_k - \frac{1}{\sqrt{A}} \ge c_k \left[1 - \frac{1}{\sqrt{A}} g(k) \right]. \tag{84}$$

From (82) $\ge$ (83), we have $a_k - c_k \ge a_j[1 - c_k g(k)]$, which is equivalent to

$$a_k - a_j \ge c_k [1 - a_j g(k)]. \tag{85}$$

Combining (84) and (85), we have

$$c_k = \min \left\{ \frac{a_k - \frac{1}{\sqrt{A}}}{1 - \frac{1}{\sqrt{A}} g(k)}, \frac{a_k - a_j}{1 - a_j g(k)}, j \ge k+1 \right\}.$$

It is not hard to verify that $a_k < \frac{1}{g(k)}$. One can verify that function

$$f(x) = \frac{a_k - x}{1 - x g(k)} = \frac{1}{g(k)} + \frac{a_k - \frac{1}{g(k)}}{1 - x g(k)}$$

is a decreasing function of x. Hence,

$$c_k = \frac{a_k - a_{k+1}}{1 - a_{k+1} g(k)}.$$

It also indicates that ϕ_{k+1} is selected in the next LARS step. This is the first result stated at the beginning of this proof. To verify (80), we need to compute the new residual:

$$\begin{aligned}
r_k &= \widetilde{r}_{k-1} - c_k u_k \\
&= s - \sum_{j=1}^{k-1} \frac{a_j - a_k}{b_j}\delta_j - c_k u_k \\
&= s - \sum_{j=1}^{k-1} \frac{a_j - a_k}{b_j}\delta_j - \frac{a_k - a_{k+1}}{1 - a_{k+1}g(k)} u_k.
\end{aligned}$$

The coefficient of s in r_k is

$$\begin{aligned}
& 1 - \frac{a_k - a_{k+1}}{1 - a_{k+1}g(k)} v_k^T \left(D_k^{-1}\mathbf{1} - D_k^{-1} v_k v_k^T D_k^{-1}\mathbf{1} \frac{1}{1 + v_k^T D_k^{-1} v_k} \right) \\
= \; & 1 - \frac{a_k - a_{k+1}}{1 - a_{k+1}g(k)} g(k) \\
= \; & \frac{1 - a_k g(k)}{1 - a_{k+1}g(k)}. \qquad (86)
\end{aligned}$$

The coefficient of δ_k in r_k is

$$\begin{aligned}
& -\frac{a_k - a_{k+1}}{1 - a_{k+1}g(k)} \cdot \frac{1}{b_k}[1 - a_k g(k)] \\
= \; & -\frac{a_k - a_{k+1}}{b_k} \cdot \frac{1 - a_k g(k)}{1 - a_{k+1}g(k)}. \qquad (87)
\end{aligned}$$

The coefficient of δ_j, $1 \le j \le k-1$, in r_k is

$$\begin{aligned}
& -\frac{a_j - a_k}{b_j} - \frac{a_k - a_{k+1}}{1 - a_{k+1}g(k)} \cdot \frac{1}{b_j}[1 - a_j g(k)] \\
= \; & -\frac{1}{b_j}\left\{ a_j - a_k + \frac{a_k - a_{k+1}}{1 - a_{k+1}g(k)}[1 - a_j g(k)] \right\} \\
= \; & -\frac{a_j - a_{k+1}}{b_j} \cdot \frac{1 - a_k g(k)}{1 - a_{k+1}g(k)}. \qquad (88)
\end{aligned}$$

Compare (86), (87) and (88), the next surrogate residual, after getting rid of factor $\frac{1 - a_k g(k)}{1 - a_{k+1}g(k)}$, is

$$\widetilde{r}_k = s - \sum_{j=1}^{k} \frac{a_j - a_{k+1}}{b_j}\delta_j.$$

This proves the second result stated at the beginning of this proof. From here, it is not hard to see that the theorem is proven. □

B.4 Proof of Theorem 4.7

For any subset of indices A, let z_A denote the subvector of z given by A, and let Φ_A denote the submatrix of Φ with the column indices in A. We only need to show that given (46), (47), and (48), function $z_A^T(\Phi_A^T\Phi_A)^{-1}z_A - F|A|$ is maximized at $A_1 = \{1, 2, ..., k\}$. In order to prove this is true, three situations are considered.

Case 1. If $|A| = k$, but A is not $\{1, 2, ..., k\}$. Recall $\Phi_1 = [\phi_1, \phi_2, ..., \phi_k]$. Let $v_1 = (z_1, z_2, ..., z_k)^T$, we have

$$v_1^T(\Phi_1^T\Phi_1)^{-1}v_1 \geq \frac{\sum_{i=1}^{k} z_i^2}{1 + (k-1)\mu} \quad .$$

On the other hand,

$$z_A^T(\Phi_A^T\Phi_A)^{-1}z_A \leq \frac{\sum_{i=1}^{k-1} z_i^2 + z_{k+1}^2}{1 - (k-1)\mu} \quad .$$

If (46) is true, recall $z_i \leq 1$, we have

$$[1 - (k-1)\mu]z_k^2 \geq 2(k-1)\mu \sum_{i=1}^{k-1} z_i^2 + z_{k+1}^2[1 + (k-1)\mu].$$

The above is equivalent to

$$[1 - (k-1)\mu] \sum_{i=1}^{k} z_i^2 \geq [1 + (k-1)\mu] \sum_{i=1}^{k-1} z_i^2 + z_{k+1}^2[1 + (k-1)\mu],$$

which is equivalent to

$$\frac{\sum_{i=1}^{k} z_i^2}{1 + (k-1)\mu} \geq \frac{\sum_{i=1}^{k-1} z_i^2 + z_{k+1}^2}{1 - (k-1)\mu} \quad .$$

Hence, we proved that

$$z_A^T(\Phi_A^T\Phi_A)^{-1}z_A \leq v_1^T(\Phi_1^T\Phi_1)^{-1}v_1.$$

This proves that A_1 maximizes $z_A^T(\Phi_A^T\Phi_A)^{-1}z_A - F|A|$ among all k-subsets.

Case 2. If $|A| > k$, assume $\ell = |A| - k$. Using a similar argument as in the previous case, one only needs to prove

$$\frac{\sum_{i=1}^{k} z_i^2}{1 + (k-1)\mu} \geq \frac{\sum_{i=1}^{k+\ell} z_i^2}{1 - (k+\ell-1)\mu} - \ell \cdot F,$$

which will guarantee that no subset with more than k variates produces a larger value of

$$z_A^T(\Phi_A^T\Phi_A)^{-1}z_A - F \cdot |A|.$$

If (47) holds, we have

$$\begin{aligned} z_{k+1}^2 &\le \sum_{i=1}^k z_i^2 \left[\frac{1-k\mu}{1+(k-1)\mu} - 1\right] = F\cdot(1-\Delta) \\ &\le \frac{1}{\ell}\sum_{i=1}^k z_i^2 \left[\frac{1-(k+\ell-1)\mu}{1+(k-1)\mu} - 1\right] + F\cdot[1-(k+\ell-1)\mu]. \end{aligned}$$

Hence,

$$\sum_{i=k+1}^{k+\ell} z_i^2 \le \sum_{i=1}^k z_i^2 \left[\frac{1-(k+\ell-1)\mu}{1+(k-1)\mu} - 1\right] + \ell\cdot F\cdot[1-(k+\ell-1)\mu].$$

Hence,

$$\frac{\sum_{i=k+1}^{k+\ell} z_i^2}{1-(k+\ell-1)\mu} \le \sum_{i=1}^k z_i^2 \left[\frac{1}{1+(k-1)\mu} - \frac{1}{1-(k+\ell-1)\mu}\right] + \ell\cdot F.$$

Hence,

$$\frac{\sum_{i=1}^k z_i^2}{1+(k-1)\mu} \ge \frac{\sum_{i=1}^{k+\ell} z_i^2}{1-(k+\ell-1)\mu} - \ell\cdot F.$$

Again, A_1 takes the maximum.

Case 3. If A has less than k variates, one only needs

$$\frac{\sum_{i=1}^k z_i^2}{1+(k-1)\mu} \ge \frac{\sum_{i=1}^{k-1} z_i^2}{1-(k-\ell-1)\mu} + \ell\cdot F. \tag{89}$$

If (48) is true, we have

$$\ell\cdot z_k^2[1+(k-1)\mu] \ge \mu(2k-\ell-2)\sum_{i=1}^k z_i^2 + \ell\cdot F[1+(k-1)\mu][1-(k-\ell-1)\mu].$$

Hence,

$$\ell\cdot z_k^2\frac{1+(k-1)\mu}{1-(k-\ell-1)\mu} \ge \frac{(2k-\ell-2)\mu}{1-(k-\ell-1)\mu}\sum_{i=1}^k z_i^2 + \ell\cdot F\cdot[1+(k-1)\mu].$$

Hence,

$$\sum_{i=k-\ell+1}^{k} z_i^2\frac{1+(k-1)\mu}{1-(k-\ell-1)\mu} \ge \sum_{i=1}^k z_i^2 \left[\frac{1+(k-1)\mu}{1-(k-\ell-1)\mu} - 1\right] + lF[1+(k-1)\mu].$$

Hence,

$$\sum_{i=1}^k z_i^2 \ge \frac{1+(k-1)\mu}{1-(k-\ell-1)\mu}\sum_{i=1}^{k-\ell} z_i^2 + lF[1+(k-1)\mu],$$

which leads to (89).

Combining the three cases, we proved that A_1 is the solution to **(P0)**. □

B.5 Proof of Theorem 4.9

Recall $v_1 = (z_1, z_2, ..., z_k)^T$. Define $v_2 = (z_{k+1}, z_{k+2}, ..., z_n)^T$. From (42), we have

$$\omega = v_2 - (\Phi_2^T \Phi_1)(\Phi_1^T \Phi_1)^{-1} \left[v_1 - \frac{\lambda}{2} \text{sign}(x_1) \right].$$

We want to show that when (49) holds, $\|\omega\|_\infty \leq \frac{\lambda}{2}$. This will imply that A_1 satisfies (42). Hence, A_1 is the minimizer in **(P0)**.

One can have for $k < j \leq n$,

$$\begin{aligned} \left\|(\phi_j^T \Phi_1)(\Phi_1^T \Phi_1)^{-1}\right\|_2 &\leq \frac{1}{1-(k-1)\mu} \left\|\phi_j^T \Phi_1\right\|_2 \\ &\leq \frac{\sqrt{k}\mu}{1-(k-1)\mu}. \end{aligned}$$

We have

$$\begin{aligned} \|\omega\|_\infty &\leq |z_{k+1}| + \max_j \left\|(\phi_j^T \Phi_1)(\Phi_1^T \Phi_1)^{-1}\right\|_2 \left\| \left(v_1 - \text{sign}(x_1) \frac{\lambda}{2} \right) \right\|_2 \\ &\leq |z_{k+1}| + \frac{\sqrt{k}\mu}{1-(k-1)\mu} \sqrt{\sum_{i=1}^{k} \left(|z_i| + \frac{\lambda}{2} \right)^2} \\ &\leq \frac{\lambda}{2}. \end{aligned}$$

A solution based on A_1 satisfies (42). Hence it is a *type-I optimal subset.* □

REFERENCES

[1] Akaike, H., "Information theory and the maximum likelihood principle," in *International Symposium on Information Theory* (Petrov, V. and Csáki, F., eds.), (Budapest), Akademiai Kiádo, 1973.

[2] Arias, E., Donoho, D. L., and Huo, X., "Near-optimal detection of geometric objects by fast multiscale methods," *IEEE Trans. Information Theory*, vol. 51, pp. 2402–2425, July 2005.

[3] Arias-Castro, E., Donoho, D. L., and Huo, X., "Adaptive multiscale detection of filamentary structures embedded in a background of Uniform random points," *Annals of Statistics*, vol. 34, February 2006.

[4] Arias-Castro, E., Donoho, D. L., Huo, X., and Tovey, C., "Connect-the-dots: how many random points can a regular curve pass through?," *Advances in Applied Probability*, vol. 37, pp. 571–603, September 2005.

[5] Atkinson, A. C., Riani, M., and Cerioli, A., *Exploring Multivariate Data with the Forward Search.* New York: Springer-Verlag, 2004. Springer series in statistics.

[6] Berlinet, A., Liese, F., and Vajda, I., "Necessary and sufficient conditions for consistency of M-estimates in regression models with general errors," *Journal of Statistical Planning and Inference*, vol. 89, pp. 243–267, August 2000.

[7] Bloomfield, P., *Fourier Analysis of Time Series: An Introduction.* John Wiley & Sons, 1976.

[8] Busch, A., Boles, W. W., Sridharan, S., and Chandran, V., "Detection of unknown forms from document images," in *Proceedings of Workshop on Digital Image Computing*, pp. 141–144, 2003.

[9] Cao, Y. and Saha, P., "Improved branch and bound method for control structure screening," *Chemical Engineering Science*, vol. 60, pp. 1555–1564, March 2005.

[10] Celebi, M. E., Aslandogan, Y. A., and Bergstresser, P. M., "Unsupervised border detection of skin lesion images," in *the Proceedings of the IEEE International Conference on Information Technology: Coding and Computing*, (Las Vegas, NV), April 2005.

[11] Chen, J. and Huo, X., "Sparse representations for multiple measurement vectors (MMV) in an over-complete dictionary," in *International Conference on Acoustic, Speech, and Signal Processing (ICASSP)*, 2005.

[12] Chen, S. S., Donoho, D. L., and Saunders, M. A., "Atomic decomposition by basis pursuit," *SIAM Rev.*, vol. 43, no. 1, pp. 129–159, 2001. Reprinted from SIAM J. Sci. Comput. 20 (1998), no. 1, 33–61.

[13] Chen, X., “An improved branch and bound algorithm for feature selection,” *Pattern Recognition Letters*, vol. 24, pp. 1925–1933, August 2003.

[14] Clyde, M. and George, E. I., “Model uncertainty,” *Statist. Sci.*, vol. 19, pp. 81–94, February 2004.

[15] Couvreur, C. and Bresler, Y., “On the optimality of the backward greedy algorithm for the subset selection problem,” *SIAM J. Matrix Anal. Appl.*, vol. 21, no. 3, pp. 797–808, 2000.

[16] Davis, G., Mallat, S., and Zhang, Z., “Adaptive time-frequency decompositions,” *Optical Engrg.*, vol. 33, pp. 2183–2191, 1994.

[17] Dodge, Y., *Statistical Data Analysis: Based on the ℓ_1-norm and Related Methods.* North-Holland, 1987.

[18] Donoho, D. and Huo, X., *Multiscale and Multiresolution Methods*, vol. 20 of *Springer Lecture Notes in Computational Science and Engineering*, ch. Beamlets and multiscale image analysis, pp. 149–196. 2002.

[19] Donoho, D. L., Elad, M., and Temlyakov, V., *Stable recovery of sparse overcomplete representations in the presense of noise.* Stanford University and University of South Carolina, 2004. Submitted manuscript.

[20] Donoho, D. L. and Huo, X., “Uncertainty principles and ideal atomic decomposition,” *IEEE Transactions on Information Theory*, vol. 47, pp. 2845–2862, November 2001.

[21] Donoho, D. L. and Huo, X., “BeamLab and reproducible research,” *International Journal of Wavelets, Multiresolution and Information Processing (IJWMIP),*, vol. 2, pp. 391–414, December 2004.

[22] Donoho, D. L. and Johnstone, I. M., “Ideal spatial adaptation by wavelet shrinkage,” *Biometrika*, vol. 81, pp. 425–455, September 1994.

[23] Donoho, D. L. and Johnstone, I. M., “Adapting to unknown smoothness via wavelet shrinkage,” *J. Amer. Statist. Assoc.*, vol. 90, no. 432, pp. 1200–1224, 1995.

[24] Efron, B., “The estimation of prediction error: covariance penalties and cross-validation,” *Journal of the American Statistical Association*, vol. 99, pp. 619–632, September 2004.

[25] Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R., “Least angle regression,” *Ann. Statist.*, vol. 32, no. 2, pp. 407–499, 2004.

[26] Ellis, S. P. and Morgenthaler, S., “Leverage and breakdown in ℓ_1 regression,” *Journal of the American Statistical Association*, vol. 87, pp. 143–148, 1992.

[27] Fan, J. and Li, R., “Variable selection via nonconvave penalized likelihood and its oracle properties,” *Journal of the American Statistical Association*, vol. 96, no. 456, pp. 1348–1360, 2001.

[28] Furnival, G. and Wilson, R., “Regression by leaps and bounds,” *Technometrics*, vol. 16, no. 4, pp. 499–511, 1974.

[29] George, E. I., "The variable selection problem," *J. Amer. Statist. Assoc.*, vol. 95, no. 452, pp. 1304–1308, 2000.

[30] George, E. I. and Foster, D. P., "The risk inflation criterion for multiple regression," *The Annals of Statistics*, vol. 22, pp. 1947–1975, 1994.

[31] George, E. L., "The variable selection problem," *Journal of the American Statistical Association*, vol. 95, pp. 1304–1308, December 2000.

[32] Gilbert, A. C., Muthukrishnan, M., and Strauss, M. J., "Approximation of functions over redundant dictionaries using coherence," in *In The 14th Annual ACM-SIAM Symposium on Discrete Algorithms*, January 2003.

[33] Gill, P. E., Murray, W., and Saunders, M. A., *User's Guide for SNOPT 5.3: A Fortran Package for Large-Scale Nonlinear Programming*, 1998. Draft.

[34] Gilmour, S. G., "The interpretation of Mallows's c_p-statistic," *Statistician*, vol. 45, no. 1, pp. 49–56, 1996.

[35] Golub, G. H. and Loan, C. F. V., *Matrix Computations.* Baltimore: Johns Hopkins University Press, 3rd ed., 1996.

[36] Gribonval, R., Figueras i Ventura, R. M., and Vandergheynst, P., "A simple test to check the optimality of sparse signal approximations," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2005. http://lts1pc19.epfl.ch/repository/Gribonval2005_1167.pdf; A longer version is available at ftp://ftp.irisa.fr/techreports/2004/PI-1661.pdf.

[37] Gribonval, R. and Nielsen, M., "Sparse representations in unions of bases," *IEEE Trans. Inform. Theory*, vol. 49, no. 12, pp. 3320–3325, 2003.

[38] Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., and Stahel, W. A., *Robust Statistics: The Approach Based on Influence Functions.* Wiley Series in Pro. and Math. Sci., 1986.

[39] Hastie, T., Rosset, S., Tibshirani, R., and Zhu, J., "The entire regularization path for the support vector machine," *Journal of Machine Learning Research*, vol. 5, pp. 1391–1415, October 2004. Also show in Neural Information Processing Systems (NIPS 2004).

[40] Hastie, T., Rosset, S., Tibshirani, R., and Zhu, J., "The entire regularization path for the Support Vector Machine," in *NIPS*, 2004.

[41] Hastie, T., Tibshirani, R., and Friedman, J., *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* New York: Springer, 2001.

[42] Huber, P. J., "Fisher information and spline interpolation," *Annals of Statistics*, vol. 2, pp. 1029–1033, 1974.

[43] Huber, P. J., *Robust Statistical Procedures*, vol. 27. CBMS-NSF, 1977.

[44] Huber, P. J., *Robust Statistics.* Wiley Series in Pro. and Math. Sci., 1981.

[45] Huo, X., “Multiscale Approximation MEthods (mame) to locate embedded consecutive subsequences – its applications in statistical data mining and spatial statistics,” *Computers & Industrial Engineering*, vol. 43, pp. 703–720, September 2002.

[46] Huo, X., *Encyclopedia of Statistical Sciences*, ch. Beamlets and multiscale modelling. N.J.: Wiley & Sons, 2 ed., 2005.

[47] Huo, X., “Exact lower bound for proportion of maximally embedded beamlet,” *Applied Mathematics Letters*, vol. 18, pp. 529–534, May 2005.

[48] Huo, X., “Minimax correlation between a line segment and a beamlet,” *Statistics & Probability Letters*, vol. 72, pp. 71–81, April 2005.

[49] Huo, X. and Chen, J., “Building a cascade detector and applications in automatic target recognition,” *Applied Optics: Information Processing (IP)*, vol. 43, pp. 293–303, January 2004.

[50] Huo, X. and Chen, J., “JBEAM: multiscale curve coding via beamlets,” *IEEE Trans. Image Processing*, vol. 14, pp. 1665–1677, November 2005.

[51] Huo, X., Donoho, D. L., Tovey, C., and Arias-Castro, E., “Dynamic programming methods for “connect the dots.”

[52] Huo, X. and Lu, J., “A network flow approach in finding maximum likelihood estimate of high concentration regions,” *Computational Statistics and Data Analysis*, vol. 46, pp. 33–56, May 2004.

[53] Huo, X. and Ni, X. S., “When do stepwise algorithms meet subset selection criteria?.” ISyE Statistics Techical Report, URL = http://www.isye.gatech.edu/statistics/papers/, July 2005.

[54] Jain, A. and Zongker, D., “Feature selection: evaluation, application, and small sample performance,” *IEEE Trans. On Pattern Analysis and Machine Intelligence*, vol. 19, pp. 153–158, February 1997.

[55] Kadane, J. B. and Lazar, N. A., “Methods and criteria for model selection,” *Journal of American Statistical Association*, vol. 99, pp. 279–290, March 2004.

[56] Kalender, W., Polacin, A., and Suess, C., “A comparison of conventional and spiral CT: an experimental study on the detection of spherical lesions,” *J. Comput. Assist. Tomogr.*, vol. 18, no. 167-176, 1994.

[57] Kim, J. R., Muller, J. P., and Morley, J., “Quantitative assessment of automated crater detection on Mars,” in *Proceedings of the XXth ISPRS Congress*, Commission IV, pp. 816–821, July 2004.

[58] Kohavi, R. and John, G. H., “Wrappers for feature subset selection,” *Artificial Intelligence*, vol. 97, pp. 273–324, December 1997.

[59] Kojadinovic, I., “Relevance measures for subset variable selection in regression problems based on k-additive mutual information,” *Computational Statistics & Data Analysis*, vol. 49, no. 4, pp. 1205–1227, 2005.

[60] Lehmann, E. L., *Theory of Point Estimation.* Pacific Grove, CA: Wadsworth & Brooks/Cole, 1991.

[61] Li, W. and Swetits, J. J., "The linear ℓ_1 estimator and the Huber M-estimator," *Siam J. Optim.*, vol. 8, no. 2, pp. 457–475, 1998.

[62] Malioutov, D. M., Cetin, M., and Willsky, A. S., "Homotopy continuation for sparse signal representation," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 5, (Philadelphia, PA), pp. 733–736, March 2005.

[63] Mallat, S., *A Wavelet Tour of Signal Processing.* San Diego, CA: Academic Press, Inc., 1998.

[64] Mallat, S. and Zhang, Z., "Matching pursuit in a time-frequency dictionary," *IEEE Trans. Signal Proc.*, vol. 41, pp. 3397–3415, 1993.

[65] Mallows, C. L., "Some comments on c_p," *Technometrics*, vol. 15, pp. 661–675, 1973.

[66] Maloof, M. A., Langley, P., Binford, T. O., Nevatia, R., and Sage, S., "Improved rooftop detection in aerial images with machine learning," *Machine Learning*, vol. 53, pp. 157–191, October-November 2003.

[67] Michelot, C. and Bougeard, M. L., "Duality results and proximal solutions of the Huber M-estimator problem," *Applied Mathematics & Optimization*, vol. 30, pp. 203–221, 1994.

[68] Miller, A. J., *Subset Selection in Regression.* New York: Chapman and Hall, 1990.

[69] Mitchell, J. S. B., Rote, G., Sundaram, G., and Woeginger, G., "Counting convex polygons in planar point sets," *Information Processing Letters*, vol. 56, no. 1, pp. 45–49, 1995.

[70] Narendra, P. M. and Fukunaga, K., "A branch and bound algorithm for feature subset selection," *IEEE Trans. Comput.*, vol. 26, pp. 917–922, September 1977.

[71] Natarajan, B. K., "Sparse approximate solutions to linear systems," *SIAM Journal on Computing*, vol. 24, no. 2, pp. 227–234, 1995.

[72] Ni, X. S. and Huo, X., "Another look at Huber's estimate in regression," 2005. Submitted manuscript.

[73] Ni, X. S. and Huo, X., "Enhanced leaps-and-bounds methods in subset selections with additional optimality tests," (San Francisco), INFORMS, November 2005. One of four finalists in the INFORMS QSR Best Student Paper Competition.

[74] Ni, X. S. and Huo, X., "On the detectability of convex-shaped inhomogeneous objects in digital images." Submitted, October 2005. Available from the authors.

[75] Ni, X. S. and Huo, X., "Regression by enhanced leaps-and-bounds via additional optimality tests (LBOT)." Submitted, 2005. Available from the authors.

[76] Noronha, S. and Nevatia, R., "Detection and description of buildings from multiple aerial images," in *Proceedings of the 1997 Conference on Computer Vision and Pattern Recognition (CVPR '97)*, pp. 588–594, 1997.

[77] Osborne, M. R., Presnell, B., and Turlach, B., "On the Lasso and its dual," *Journal of Computational and Graphical Statistics*, vol. 9, no. 2, pp. 319–337, 2000.

[78] Osborne, M. R., Presnell, B., and Turlach, B. A., "A new approach to variable selection in least squares problems," *IMA J. Numer. Anal.*, vol. 20, no. 3, pp. 389–403, 2000.

[79] Pati, Y. C., Rezaiifar, R., and Krishnaprasad, P. S., "Orthogonal matching pursuit: recursive function approximation with applications to wavelet decomposition," in *Proc. 27th Asilomar Conference on Signals, Systems and Computers* (Singh, A., ed.), (Los Alamitos, CA), IEEE Comput. Soc. Press, 1993.

[80] Pollard, D., *Convergence of Stochastic Processes.* New York: Springer-Verlag, 1984.

[81] Qi, J., "Analysis of lesion detectability in Bayesian emission reconstruction with nonstationary object variability," *IEEE Transactions on Medical Imaging*, vol. 23, no. 3, pp. 321–329, 2004.

[82] Qi, J. and Huesman, R. H., "Theoretical study of lesion detectability of MAP reconstruction using computer observers," *IEEE Trans. Med. Imaging*, vol. 20, pp. 815–822, August 2001.

[83] Rockafellar, R. T., *Convex Analysis.* Princeton, NJ: Princeton University Press, 1970.

[84] Schwarz, G., "Estimating the dimension of a model," *The Annals of Statistics*, vol. 6, pp. 461–464, 1978.

[85] Shen, X. T., Huang, H. C., and Ye, J., "Inference after model selection," *Journal of the American Statistical Association*, vol. 99, pp. 751–762, September 2004.

[86] Shen, X. T. and Ye, J. M., "Adaptive model selection," *Journal of the American Statistical Association*, vol. 97, pp. 210–221, March 2002.

[87] Silva, A. P. D., "Efficient variable screening for multivariate analysis," *Journal of Multivariate Analysis*, vol. 76, pp. 35–62, January 2001.

[88] Somol, P., Pudil, P., and Kittler, J., "Fast branch & bound algorithms for optimal feature selection," *IEEE Trans. On Pattern Analysis and Machine Intelligence*, vol. 26, pp. 900–912, July 2004.

[89] Spingarn, J. E., "Partial inverse of a monotone operator," *Applied Mathematics and Optimization*, vol. 10, pp. 247–265, 1983.

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

[91] Tropp, J. A., "Greed is good: Algorithmic results for sparse approximation," tech. rep., CAM Tech Report, Univ. Texas, 2003.

[92] Tropp, J. A., "Just relax: Convex programming methods for subset selection and sparse approximation," tech. rep., ICES Report 04-04,UT-Austin, 2004.

[93] van Heel, M., Gowen, B., Matadeen, R., Orlova, E. V., Finn, R., Pape, T., Cohen, D., Stark, H., Schmidt, R., Schatz, M., and Patwardhan, A., "Single-particle electron cryo-microscopy: towards atomic resolution," *Quarterly Reviews of Biophysics*, vol. 33, pp. 307–369, November 2000.

[94] Vanderbei, R. J., *Linear Programming.* Kluwer Academic Publishers, 1996.

[95] Weisberg, S., "Discussion of [25]," *The Annals of Statistics*, vol. 32, no. 2, pp. 490–494, 2004.

[96] Wu, C. F. J., "Construction of supersaturated designs through partially aliased interactions," *Biometrika*, vol. 80, pp. 661–669, September 1993.

[97] Xu, L., Jackowski, M., Goshtasby, A., Yu, C., Roseman, D., Bines, S., Dhawan, A., and Huntley, A., "Segmentation of skin cancer images," *Image and Vision Computing*, vol. 17, no. 1, pp. 65–74, 1999.

[98] Ye, J. M., "On measuring and correcting the effects of data mining and model selection," *Journal of the American Statistical Association*, vol. 93, pp. 120–131, March 1998.

[99] Yu, B. and Yuan, B., "A more efficient branch and bound algorithm for feature selection," *Pattern Recognition*, vol. 26, pp. 883–889, June 1993.

[100] Yu, Z. 'Particle Picking' website. http://www.ices.utexas.edu/ zeyun/pick.htm.

[101] Zheng, X. D. and Loh, W. Y., "Consistent variable selection in linear models," *Journal of the American Statistical Association*, vol. 90, pp. 151–156, March 1995.

[102] Zhu, Y., Carragher, B., and Potter, C. S., "Contaminant detection: improving template matching based particle selection for cryo-electron microscopy," in *Proceedings of the IEEE International Symposium on Biomedical Imaging*, (Arlington VA), pp. 1071–1074, April 2004.

[103] Zou, H., Hastie, T., and Tibshirani, R., "On the "degrees of freedom" of the Lasso." Submitted manuscript. Available at http://www-stat.stanford.edu/∼hastie/Papers/, 2004.