การเปรียบเทียบประสิทธิภาพของการประมาณค่าของพารามิเตอร์ด้วยวิธีลาสโซและวิธีการคัดเลือกชุดข้อมูลย่อยที่ดีที่สุดในการวิเคราะห์การถดถอยเชิงเส้นสำหรับข้อมูลที่มีมิติสูง

น.ส.วรัญญา บุตรบุรี

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
สาขาวิชาสถิติ ภาควิชาสถิติ
คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2564
ลิขสิทธิ์ของจุฬาลงกรณ์มหาวิทยาลัย

A PERFORMANCE COMPARISON OF PARAMETER ESTIMATIONS BY LASSO METHOD AND THE BEST SUBSET SELECTION METHOD IN LINEAR REGRESSION ANALYSIS FOR HIGH DIMENSIONAL DATA

Miss Waranya Butburee

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Science in Statistics
Department of Statistics
FACULTY OF COMMERCE AND ACCOUNTANCY
Chulalongkorn University
Academic Year 2021

Copyright of Chulalongkorn University

หัวข้อวิทยานิพนธ์	การเปรียบเทียบประสิทธิภาพของการประมาณค่าของพารามิเตอร์ด้วยวิธีลาสโซและวิธีการคัดเลือกชุดข้อมูลย่อยที่ดีที่สุดในการวิเคราะห์การถดถอยเชิงเส้นสำหรับข้อมูลที่มีมิติสูง
โดย	น.ส.วรัญญา บุตรบุรี
สาขาวิชา	สถิติ
อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก	รองศาสตราจารย์ ดร.เสกสรร เกียรติสุไพบูลย์

คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต

..... คณบดีคณะพาณิชยศาสตร์และการบัญชี

(รองศาสตราจารย์ ดร.วิเลิศ ภูริวัชร)

คณะกรรมการสอบวิทยานิพนธ์

..... ประธานกรรมการ

(ผู้ช่วยศาสตราจารย์ ดร.วิฐรา พึ่งพาพงศ์)

..... อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก

(รองศาสตราจารย์ ดร.เสกสรร เกียรติสุไพบูลย์)

..... กรรมการ

(ผู้ช่วยศาสตราจารย์ ดร.สุรณพีร์ ภูมิวุฒิสาร)

..... กรรมการภายนอกมหาวิทยาลัย

(รองศาสตราจารย์ ดร.สันติ ถิรพัฒน์)

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

วรัญญา บุตรบุรี : การเปรียบเทียบประสิทธิภาพของการประมาณค่าของพารามิเตอร์ด้วยวิธีลาสโซและวิธีการคัดเลือกชุดข้อมูลย่อยที่ดีที่สุดในการวิเคราะห์การถดถอยเชิงเส้นสำหรับข้อมูลที่มีมิติสูง. (A PERFORMANCE COMPARISON OF PARAMETER ESTIMATIONS BY LASSO METHOD AND THE BEST SUBSET SELECTION METHOD IN LINEAR REGRESSION ANALYSIS FOR HIGH DIMENSIONAL DATA) อ.ที่ปรึกษาหลัก : รศ. ดร.เสกสรร เกียรติสุไพบูลย์

งานวิจัยครั้งนี้มีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าพารามิเตอร์สำหรับข้อมูลที่มีมิติสูงด้วยทั้งหมด 5 วิธี ได้แก่ วิธี L0Learn, L0L2Learn, L1, A-L1 และวิธี A-L1L2 โดยการเปรียบเทียบประสิทธิภาพจะเปรียบเทียบใน 2 ด้าน คือ 1) เปรียบเทียบประสิทธิภาพด้านการพยากรณ์ ซึ่งวัดจากค่าคลาดเคลื่อนการทำนาย (MSE) และ 2) ความถูกต้องในการคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบ ซึ่งพิจารณาจากของค่า Precision Recall และค่า AUC ข้อมูลที่มีมิติสูงที่ใช้ในการศึกษาครั้งนี้ได้จากการจำลอง โดยกำหนดให้ในแต่ละชุดข้อมูลประกอบด้วยจำนวนค่าสังเกต 100 ค่าสังเกต (n = 100) และมีตัวแปรอิสระจำนวน 100 ตัว (p = 1000) โดยตัวแปรอิสระมีการแจกแจงแบบปรกติหลายตัวแปรซึ่งมีความสัมพันธ์กันแบบยกกำลัง (Exponential Correlation) 3 ระดับคือ 0, 0.5 และ 0.9 ค่าความคลาดเคลื่อนสุ่มขึ้นอยู่กับอัตราส่วนสัญญาณต่อสัญญาณรบกวน (SNR) ซึ่งมี 6 ระดับคือ 0.1, 0.5, 1, 5, 10, และ 20 โดยจำลองข้อมูลจำนวน 100 ชุดในแต่ละสถานการณ์ จากการวัดประสิทธิภาพจากค่าเฉลี่ยของข้อมูลทั้ง 100 ชุด ผลการเปรียบเทียบประสิทธิภาพด้านการพยากรณ์พบว่า เมื่อข้อมูลมีค่า SNR ต่ำและตัวแปรอิสระมีความสัมพันธ์กันน้อยถึงปานกลาง วิธี L1 จะมีประสิทธิภาพสูงที่สุด ตามด้วยวิธี L0L2Leran วิธี L0Learn วิธี A-L1L2 และวิธี A-L1 ตามลำดับ แต่เมื่อข้อมูลมีค่า SNR เพิ่มสูงขึ้นและในขณะเดียวกันตัวแปรอิสระมีความสัมพันธ์กันมากขึ้นวิธี A-L1 และวิธี A-L1L2 จะมีประสิทธิภาพสูงที่สุด ตามด้วยวิธี L1 วิธี L0L2Leran วิธี L0Learn ตามลำดับ ส่วนผลการเปรียบเทียบประสิทธิภาพด้านการคัดเลือกตัวแปรเข้าสู่ตัวแบบ เมื่อพิจารณาจากค่าเฉลี่ยของค่า Precision วิธี L0Learn และวิธี L0L2Learn มีประสิทธิภาพมากกว่าวิธีอื่น ๆ และเมื่อพิจารณาจากค่าเฉลี่ยของค่า Recall ในกรณีข้อมูลมีค่า SNR ต่ำวิธี A-L1 และวิธี A-L1L2 จะมีประสิทธิภาพมากที่สุด รองลงมาคือวิธี L0L2Learn วิธี L1 และวิธี L0Learn ตามลำดับ แต่เมื่อข้อมูลมีค่า SNR มากขึ้นและตัวแปรอิสระมีความสัมพันธ์กันมากขึ้น วิธี L1 มีประสิทธิภาพสูงที่สุดเทียบเท่ากับวิธี A-L1 และวิธี A-L1L2 และเมื่อพิจารณาจากค่าเฉลี่ยของค่า AUC กรณีข้อมูลมีค่า SNR ต่ำและตัวแปรอิสระมีความสัมพันธ์กันน้อย วิธี L0L2Learn วิธี L1 วิธี A-L1 และวิธี A-L1L2 จะมีประสิทธิภาพใกล้เคียงกันและมีประสิทธิภาพมากกว่าวิธี L0Learn แต่เมื่อข้อมูลมีค่า SNR มากขึ้นและตัวแปรอิสระมีความสัมพันธ์กันมากขึ้นวิธี L0L2Learn และวิธี A-L1L2 จะมีประสิทธิภาพดีกว่าวิธีอื่น ๆ นอกจากนี้ยังพบว่าวิธี L0Learn และวิธี L0L2Learn จะให้ตัวแบบที่มีขนาดเล็กส่งผลให้ตัวแบบมีค่า Precision โดยเฉลี่ยสูง และมีข้อดีคือตัวแบบอธิบายได้ง่าย ในทางตรงกันข้ามวิธี L1 วิธี A-L1 และวิธี A-L1L2 จะให้ตัวแบบที่มีขนาดใหญ่กว่าส่งผลให้มีค่า Recall โดยเฉลี่ยสูง แต่มีข้อจำกัดคือตัวแบบอธิบายได้ยาก

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

สาขาวิชา	สถิติ	ลายมือชื่อนิสิต .....
ปีการศึกษา	2564	ลายมือชื่อ อ.ที่ปรึกษาหลัก .....

6280284826 : MAJOR STATISTICS

KEYWORD: BEST SUBSET SELECSION, LASSO REGRESSION, SPARSE REGRESSION, REGULARIZED REGRESSION, HIGH-DIMENSIONAL DATA

Waranya Butburee : A PERFORMANCE COMPARISON OF PARAMETER ESTIMATIONS BY LASSO METHOD AND THE BEST SUBSET SELECTION METHOD IN LINEAR REGRESSION ANALYSIS FOR HIGH DIMENSIONAL DATA. Advisor: Assoc. Prof. SEKSAN KIATSUPAIBUL, Ph.D.

The objective of this research is to compare the performances of parameter estimations from five sparse regression methods for high-dimensional data, namely L0Learn, L0L2Learn, L1, A-L1 and A-L1L2. The models compared based on two criterions: 1) the predictive performance measured by the prediction mean square error and 2) the variable selection accuracy measured by the variable selection precision, the variable selection recall, and the variable selection AUC. The high-dimensional data sets under this study are simulated data sets. Each data set contains one hundred observations (n = 100) and has one thousand predictors (p = 1000). The predictors are generated based on the multivariate normal distribution whose exponential correlation parameters are set at three levels: 0.0, 0.5 and 0.9. The random error terms are simulated to possess six levels of signal-to-noise ratio (SNR): 0.1, 0.5, 1, 5, 10 and 20. At each pair of correlation parameter and SNR, one hundred data sets are generated. The performances of each method are averaged over these one hundred data sets. The results show that, in terms of the predictive performance, at low SNR and low correlation, L1 performs best followed by L0L2Leran, L0Learn, A-L1L2 and A-L1, respectively. At high SNR and high correlation, A-L1 and A-L1L2 perform best followed by L1, L0L2Learn and L0Learn, respectively. In terms of the variable selection accuracy, when measured by the variable selection precision, overall L0Learn and L0L2Learn perform significantly better than L1, A-L1, and A-L1L2. When measured by the variable selection recall, at low SNR, A-L1 and A-L1L2 perform best followed by L0L2Learn, L1 and L0Learn respectively. At high SNR and high correlation, the performances of L1, A-L1 and A-L1L2 are close to one another, but significantly better than that of L0Learn and L0L2Learn. When measured by the variable selection AUC, at low SNR and low correlation, the performances of L0L2Learn, L1, A-L1 and A-L1L2 are close to one another, but significantly better than that of L0Learn. With variable selection AUC, at high SNR and high correlation, L0L2Learn and A-L1L2 performs significantly better than other methods. Overall, L0Learn and L0L2Leaarn select a small number of variables into the model, leading to high variable selection precision. In contrast, L1, A-L1 and A-L1L2 select a larger number of variables into the model, resulting in high variable selection recall.

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

Field of Study:	Statistics	Student's Signature .....
Academic Year:	2021	Advisor's Signature .....

กิตติกรรมประกาศ

การดำเนินงานวิจัยเรื่อง "การเปรียบเทียบประสิทธิภาพของการประมาณค่าของพารามิเตอร์ด้วยวิธีลาสโซและวิธีการคัดเลือกชุดข้อมูลย่อยที่ดีที่สุดในการวิเคราะห์การถดถอยเชิงเส้นสำหรับข้อมูลที่มีมิติสูง" นี้สำเร็จลุล่วงได้เป็นอย่างดีเนื่องจากได้รับความอนุเคราะห์จากรองศาสตราจารย์ ดร. เสกสรร เกียรติสุไพบูลย์ อาจารย์ที่ปรึกษาวิทยานิพนธ์ ที่กรุณาให้คำแนะนำ คำปรึกษา ตรวจทานและแก้ไขข้อบกพร่องต่าง ๆ ให้วิทยานิพนธ์มีความครบถ้วนสมบูรณ์มากขึ้น ตลอดจนเป็นกำลังใจและแรงขับเคลื่อนให้วิทยานิพนธ์เล่มนี้ ผู้วิจัยจึงขอขอบพระคุณอาจารย์เป็นอย่างสูงไว้ ณ ที่นี้

ขอขอบพระคุณผู้ช่วยศาสตราจารย์ ดร. วิฐรา พึ่งพาพงศ์ ในฐานะเป็นผู้ให้ข้อมูลสำหรับการศึกษาต้นแบบ ให้คำปรึกษาและข้อชี้แนะในการทำวิทยานิพนธ์เสมอมา

ขอขอบพระคุณผู้ช่วยศาสตราจารย์ ดร. วิฐรา พึ่งพาพงศ์ ที่กรุณาเป็นประธานกรรมการ รองศาสตราจารย์ ดร. สันติ ถิรพัฒน์ และผู้ช่วยศาสตราจารย์ ดร. สุรณพีร์ ภูมิวุฒิสาร ที่กรุณาร่วมเป็นกรรมการในการสอบวิทยานิพนธ์ครั้งนี้ ตลอดจนให้ความกรุณาในการตรวจทาน ให้คำแนะนำและข้อชี้แนะที่เป็นประโยชน์ ทำให้วิทยานิพนธ์ฉบับนี้มีความสมบูรณ์มากขึ้น

ขอบคุณเพื่อน ๆ ภาควิชาสถิติ รุ่นพี่ รุ่นน้องทุกท่าน ที่ให้คำแนะนำ ให้ความช่วยเหลือและให้กำลังใจ สุดท้ายนี้ขอขอบคุณครอบครัวที่เป็นแรงสนับสนุน และเป็นกำลังใจที่ยิ่งใหญ่สำหรับผู้ทำวิจัยเสมอมา

วรัญญา บุตรบุรี

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

สารบัญ

หน้า

..... ค

บทคัดย่อภาษาไทย ..... ค

..... ง

บทคัดย่อภาษาอังกฤษ ..... ง

กิตติกรรมประกาศ ..... จ

สารบัญ ..... ฉ

สารบัญตาราง ..... ฌ

สารบัญรูปภาพ ..... ญ

บทที่ 1 บทนำ 1

1.1 ที่มาและความสำคัญ 1

1.2 วัตถุประสงค์การวิจัย 4

1.4 ขอบเขตของงานวิจัย 4

1.5 เกณฑ์การวัดประสิทธิภาพที่ใช้ในการวิจัย 5

1.6 ประโยชน์ที่คาดว่าจะได้รับ 6

1.7 คำจำกัดความที่ใช้ในงานวิจัย 6

บทที่ 2 ทฤษฎี แนวคิดและงานวิจัยที่เกี่ยวข้อง 8

2.1 ตัวแบบการถดถอยเชิงเส้นพหุคูณ (Multiple linear regression model) 8

2.2 การวิเคราะห์การถดถอยเชิงเส้นด้วยวิธีกำลังสองน้อยที่สุด (Least Squares Method) 9

2.2 การวิเคราะห์การถดถอยเชิงเส้นด้วยฟังก์ชันลงโทษ (Penalized Regression) 11

2.2.1 การถดถอยเชิงเส้นพหุคูณด้วยวิธีการคัดเลือกชุดข้อมูลย่อยที่ดีที่สุด (Best Subsets Selection) 12

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

2.2.2 การถดถอยเชิงเส้นพหุคูณด้วยวิธีลาสโซ (Least Absolute Selection and Shrinkage Operator)..... 17
2.2.3 การถดถอยเชิงเส้นพหุคูณด้วยวิธีลาสโซปรับได้ (Adaptive Lasso Regression)..... 19
2.2.4 การถดถอยเชิงเส้นพหุคูณด้วยวิธีอิลาสติคเน็ตแบบปรับได้ (Adaptive Elastic Net Regression Method, Adaptive Elastic Net)..... 21
2.3 การเลือกพารามิเตอร์การปรับแต่ง (Tuning Parameter) ด้วยวิธีการตรวจสอบไขว้ (Cross Validation)..... 23
2.4 งานวิจัยในอดีตที่เกี่ยวข้อง..... 25
บทที่ 3 วิธีดำเนินการวิจัย..... 31
3.1 ขอบเขตของการวิจัยและการจำลองข้อมูล..... 31
3.2 เกณฑ์ที่ใช้ในการพิจารณา..... 32
3.2.1. ประสิทธิภาพของการพยากรณ์..... 32
3.2.2. ประสิทธิภาพในการเลือกตัวประมาณ..... 33
3.3 วิธีการดำเนินงานวิจัย..... 36
3.3.1 ขั้นตอนการสร้างข้อมูล..... 36
3.3.2 ขั้นตอนการหาตัวประมาณด้วยวิธี L0Learn..... 36
3.3.3 ขั้นตอนการหาตัวประมาณด้วยวิธี L0L2Learn..... 37
3.3.4 ขั้นตอนการหาตัวประมาณด้วยวิธี L1..... 39
3.3.5 ขั้นตอนการหาตัวประมาณด้วยวิธี A-L1..... 40
3.3.6 วิธีการดำเนินงานวิจัยด้วยวิธี A-L1L2..... 41
3.4 แผนภาพการทำงานของโปรแกรม..... 43
บทที่ 4 ผลการวิจัย..... 44
4.1 ประสิทธิภาพด้านการพยากรณ์..... 44
4.2 ประสิทธิภาพด้านการคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบ..... 49
4.2.1 ค่าเฉลี่ยของค่าความแม่นยำ (Average of Precision)..... 49

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

4.2.2 ค่าเฉลี่ยของค่าความระลึก (Average of Recall) 54
4.2.3 ค่าเฉลี่ยของค่าพื้นที่ใต้กราฟ ROC (Average of AUC Score) 59
บทที่ 5 สรุปและอภิปรายผลการวิจัย 64
5.1 สรุปผลการวิจัย 64
5.2 อภิปรายผลการวิจัย 68
5.3 ข้อจำกัดของงานวิจัย และข้อเสนอแนะ 70
บรรณานุกรม 71
ประวัติผู้เขียน 76

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

สารบัญตาราง

หน้า

ตารางที่ 2.1 งานวิจัยในอดีตที่เกี่ยวข้อง 25

ตารางที่ 4.1 ค่าเฉลี่ยของค่าความคลาดเคลื่อนกำลังสอง (AMSE) จาก 100 ชุดข้อมูล เมื่อตัวแปรอิสระไม่มีความสัมพันธ์กัน 45

ตารางที่ 4.2 ค่าเฉลี่ยของค่าความคลาดเคลื่อนกำลังสอง (AMSE) จาก 100 ชุดข้อมูล เมื่อตัวแปรอิสระมีความสัมพันธ์กันปานกลาง 47

ตารางที่ 4.3 ค่าเฉลี่ยของค่าความคลาดเคลื่อนกำลังสอง (AMSE) จาก 100 ชุดข้อมูล เมื่อตัวแปรอิสระมีความสัมพันธ์กันสูง 48

ตารางที่ 4.4 ค่าเฉลี่ยของค่า Precision จาก 100 ชุดข้อมูล เมื่อตัวแปรอิสระไม่มีความสัมพันธ์กัน 50

ตารางที่ 4.5 ค่าเฉลี่ยของค่า Precision จาก 100 ชุดข้อมูล เมื่อตัวแปรอิสระมีความสัมพันธ์กันปานกลาง 52

ตารางที่ 4.6 ค่าเฉลี่ยของค่า Precision จาก 100 ชุดข้อมูล เมื่อตัวแปรอิสระมีความสัมพันธ์กันสูง 53

ตารางที่ 4.7 ค่าเฉลี่ยของค่า Recall จาก 100 ชุดข้อมูล เมื่อตัวแปรอิสระไม่มีความสัมพันธ์กัน 55

ตารางที่ 4.8 ค่าเฉลี่ยของค่า Recall จาก 100 ชุดข้อมูล เมื่อตัวแปรอิสระมีความสัมพันธ์กันปานกลาง 56

ตารางที่ 4.9 ค่าเฉลี่ยของค่า Recall จาก 100 ชุดข้อมูล เมื่อตัวแปรอิสระมีความสัมพันธ์กันสูง 58

ตารางที่ 4.10 ค่าเฉลี่ยของค่า AUC Score จาก 100 ชุดข้อมูล เมื่อตัวแปรอิสระไม่มีความสัมพันธ์กัน 60

ตารางที่ 4.11 ค่าเฉลี่ยของค่า AUC Score จาก 100 ชุดข้อมูล เมื่อตัวแปรอิสระมีความสัมพันธ์กันปานกลาง 61

ตารางที่ 4.12 ค่าเฉลี่ยของค่า AUC Score จาก 100 ชุดข้อมูล เมื่อตัวแปรอิสระมีความสัมพันธ์กันสูง 63

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

สารบัญรูปภาพ

	หน้า
รูปที่ 3.1 การคำนวณพื้นที่ใต้กราฟโดยใช้หลักสี่เหลี่ยมคางหมู (Baynes et al., 2012)	35
รูปที่ 3.2 แผนภาพการทำงานของโปรแกรม	43
รูปที่ 4.1 Boxplot แสดงค่าความคลาดเคลื่อนกำลังสอง (MSE) เมื่อตัวแปรอิสระไม่มีความสัมพันธ์กัน	45
รูปที่ 4.2 4.2 Boxplot แสดงค่าความคลาดเคลื่อนกำลังสอง (MSE) เมื่อตัวแปรอิสระมีความสัมพันธ์กันปานกลาง	47
รูปที่ 4.3 Boxplot แสดงค่าความคลาดเคลื่อนกำลังสอง (MSE) เมื่อตัวแปรอิสระมีความสัมพันธ์กันสูง	49
รูปที่ 4.4 Boxplot แสดงค่า Precision เมื่อตัวแปรอิสระไม่มีความสัมพันธ์กัน	51
รูปที่ 4.5 Boxplot แสดงค่า Precision เมื่อตัวแปรอิสระมีความสัมพันธ์กันปานกลาง	52
รูปที่ 4.6 Boxplot แสดงค่า Precision เมื่อตัวแปรอิสระมีความสัมพันธ์กันปานสูง	54
รูปที่ 4.7 Boxplot แสดงค่า Recall เมื่อตัวแปรอิสระมีไม่ความสัมพันธ์กัน	55
รูปที่ 4.8 Boxplot แสดงค่า Recall เมื่อตัวแปรอิสระมีความสัมพันธ์กันปานกลาง	57
รูปที่ 4.9 Boxplot แสดงค่า Recall เมื่อตัวแปรอิสระมีความสัมพันธ์กันสูง	59
รูปที่ 4.10 Boxplot แสดงค่า AUC Score เมื่อตัวแปรอิสระไม่มีความสัมพันธ์กัน	60
รูปที่ 4.11 Boxplot แสดงค่า AUC Score เมื่อตัวแปรอิสระมีความสัมพันธ์กันปานกลาง	62
รูปที่ 4.12 Boxplot แสดงค่า AUC Score เมื่อตัวแปรอิสระมีความสัมพันธ์กันสูง	63
รูปที่ 5.1 ค่า Log ของค่าเฉลี่ยค่าความคลาดเคลื่อนกำลังสอง (log(AMSE)) เมื่อข้อมูลมีค่า SNR และตัวแปรอิสระมีความสัมพันธ์กันในระดับแตกต่างกัน	67
รูปที่ 5.2 ค่าเฉลี่ยของค่า Precision เมื่อข้อมูลมีค่า SNR และตัวแปรอิสระมีความสัมพันธ์กันในระดับแตกต่างกัน	67
รูปที่ 5.3 ค่าเฉลี่ยของค่า Recall เมื่อข้อมูลมีค่า SNR และตัวแปรอิสระมีความสัมพันธ์กันในระดับแตกต่างกัน	67
รูปที่ 5.2 ค่าเฉลี่ยของค่าพื้นที่ใต้กราฟ ROC (Average of AUC Score) เมื่อข้อมูลมีค่า SNR และตัวแปรอิสระมีความสัมพันธ์กันในระดับแตกต่างกัน	68
รูปที่ 5.3 ค่าเฉลี่ยของจำนวนตัวแปรอิสระที่ถูกคัดเลือกเข้าสู่ตัวแบบ เมื่อข้อมูลมีค่า SNR และตัวแปรอิสระมีความสัมพันธ์กันในระดับแตกต่างกัน	68

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

บทที่ 1
บทนำ

1.1 ที่มาและความสำคัญ

การวิเคราะห์การถดถอยเชิงเส้น (Linear Regression) เป็นวิธีการวิเคราะห์และอธิบายความสัมพันธ์ระหว่างตัวแปรอิสระ (Independent Variable) กับตัวแปรตาม (Dependent Variable) ในรูปแบบความสัมพันธ์เชิงเส้นตรง วิธีถูกพัฒนาขึ้นมาตั้งแต่ในยุคสมัยก่อนที่จะใช้คอมพิวเตอร์ในทางสถิติ (Pre-computer Age of Statistics) ด้วยซ้ำไป จนมาถึงปัจจุบันซึ่งถือเป็นยุคคอมพิวเตอร์ (Computer Era) วิธี Linear Regression ก็ยังคงเป็นวิธีที่ได้รับความนิยม เนื่องจากเป็นตัวแบบที่เข้าใจได้ง่าย สามารถอธิบายผลกระทบของตัวแปรต้นที่นำเข้าสู่ตัวแบบต่อตัวแปรตามได้อย่างเพียงพอและตรงไปตรงมา นอกจากนี้ในบางกรณียังมีประสิทธิภาพในการพยากรณ์ค่าของตัวแปรตามได้ดีกว่าการวิเคราะห์การถดถอยที่ไม่เป็นเชิงเส้นตรง (Nonlinear Models) โดยเฉพาะในกรณีที่ข้อมูลมีจำนวนตัวอย่างน้อย และมีอัตราส่วนสัญญาณต่อสัญญาณรบกวน (Signal to Noise Ratio: SNR) ต่ำ (Trevor Hastie et al., 2017)

การวิเคราะห์การถดถอยเชิงเส้นสามารถใช้ในการวิเคราะห์ได้อย่างหลากหลาย ไม่ว่าจะต้องการวิเคราะห์ความสัมพันธ์ของตัวแปรต้นและตัวแปรตามเพียงอย่างหนึ่งตัว หรือที่เรียกกันว่าการวิเคราะห์การถดถอยเชิงเส้นอย่างง่าย (Simple Linear Regression) หรือการวิเคราะห์การถดถอยเชิงเส้นพหุคูณ (Multiple Linear Regression) ที่ใช้ในการศึกษาความสัมพันธ์ของตัวแปรตาม 1 ตัวกับตัวแปรอิสระที่มีตั้งแต่ 2 ตัวขึ้นไป และข้อมูลมีขนาดตัวอย่าง (n) มากกว่าจำนวนตัวแปรอิสระ (p) ซึ่งเรียกว่าข้อมูลที่มีมิติต่ำ (Low Dimensional Data) โดยวิธีการประมาณค่าที่นิยมใช้คือวิธีกำลังสองน้อยที่สุด (Ordinary Least Squares Method: OLS) ซึ่งตัวประมาณที่ได้จะมีคุณสมบัติของตัวประมาณที่ดีคือ ไม่เอนเอียง (Unbiased) มีความแปรปรวนต่ำ (Minimum Variance) มีความคงเส้นคงวา (Consistency) มีความพอเพียง (Sufficiency) และมีประสิทธิภาพสูงสุด (Efficiency)

ในปัจจุบันคอมพิวเตอร์และเทคโนโลยีต่าง ๆ ได้รับการพัฒนาอย่างรวดเร็ว ข้อมูลที่เกิดขึ้นมีจำนวนมาก และมีความซับซ้อนมากขึ้น นอกจากนี้ความก้าวหน้าทางเทคโนโลยีในปัจจุบันช่วยให้เราสามารถจัดเก็บข้อมูลขนาดใหญ่ในรูปแบบอิเล็กทรอนิกส์ได้อย่างมีประสิทธิภาพ ข้อมูลบางประเภทมีตัวแปรจำนวนมากและบางครั้งอาจมีมากกว่าจำนวนตัวอย่าง ซึ่งเรียกข้อมูลที่มีจำนวนตัวแปรอิสระ (p) มากกว่าขนาดตัวอย่าง (n) ว่าข้อมูลที่มีมิติสูง (High dimensional Data) (Hastie et al., 2015) ซึ่งข้อมูลที่มีมิติสูงมักพบได้มากในข้อมูลทางการแพทย์ โดยเฉพาะข้อมูลที่มีความเกี่ยวข้องกับ DNA หรือ RNA เช่น ข้อมูลไมโครอาร์เรย์ (Microarray) ของ DNA เนื้องอกในเต้านม โดยตัวแปรอิสระ

คือระดับการแสดงออกของยีน (Gene Expression) กว่า 20,000 ยีน แต่มีจำนวนตัวอย่างที่รวบรวมได้เพียง 377 คนเท่านั้น (Van't Veer et al., 2002) เป็นต้น ในกรณีที่มีตัวแปรอิสระเป็นจำนวนมาก อาจจะมีตัวแปรอิสระบางตัวเท่านั้นที่ควรอยู่ในตัวแบบ เรียกว่า ตัวแปรที่มีผล (Active Variable) และบางตัวไม่ควรอยู่ในตัวแบบ (Inactive Variable) นั่นคือ ค่าสัมประสิทธิ์การถดถอยมีค่าเท่ากับศูนย์ และถ้าในตัวแบบมีค่าสัมประสิทธิ์การถดถอยส่วนใหญ่เป็นศูนย์ จะเรียกลักษณะตัวแบบนี้ว่า ตัวแบบบางเบา (Sparse Model) (ชุติกาญจน์ ชูสวัสดิ์, 2560)

การวิเคราะห์การถดถอยเชิงเส้น (Linear Regression) ไม่สามารถวิเคราะห์ข้อมูลที่มีมิติสูงได้ เนื่องจากการวิเคราะห์การถดถอยเชิงเส้นมีเงื่อนไขว่าจำนวนตัวอย่างจะต้องมีจำนวนมากกว่าจำนวนตัวแปรอิสระ (วิฐรา พึ่งพาพงศ์, 2558) หากนำข้อมูลที่มีมิติสูงไปใช้การวิเคราะห์การถดถอยเชิงเส้นปัญหาที่มักจะพบบ่อยคือ 1) ไม่สามารถประมาณค่าสัมประสิทธิ์การถดถอยได้ด้วยวิธีกำลังสองน้อยที่สุด 2) ตัวแปรอิสระมีโอกาสที่จะมีความสัมพันธ์เชิงเส้นกัน หรือเกิดความสัมพันธ์เชิงเส้นแบบพหุ (Multicollinearity) และ 3) มีโอกาสที่ตัวแปรตามจะมีความสัมพันธ์กับตัวแปรอิสระเป็นจำนวนมาก ตัวแบบที่ได้มีขนาดใหญ่ทำให้ยากต่อการแปรผล (พัชราภรณ์ พรดำเนินสวัสดิ์, 2560; วิฐรา พึ่งพาพงศ์, 2558)

การวิเคราะห์ข้อมูลที่มีมิติสูงจะนิยมใช้วิธีการถดถอยเชิงเส้นแบบพีนอลไลซ์ (Penalized Regression) เป็นวิธีที่เพิ่มฟังก์ชันการลงโทษ (Penalty Function) เข้าไปในสมการที่ใช้ประมาณค่าสัมประสิทธิ์การถดถอย วิธี Penalized Regression วิธี Least Absolute Selection and Shrinkage Operator (Lasso ซึ่งในงานวิจัยนี้เรียกว่า L1) ถูกนำเสนอโดย Tibshirani ในปี1996 (Tibshirani, 1996) L1 ใช้ L_1norm Penalty Function (ฟังก์ชันการลงโทษด้วยผลรวมขนาดพารามิเตอร์ที่ถูกถ่วงน้ำหนัก) ในการประมาณสัมประสิทธิ์การถดถอย เป็นวิธีหนึ่งที่นิยมใช้กันในวงกว้างในปัจจุบัน เพราะสามารถคัดเลือกตัวแปรโดยการหดขนาด (Shrinkage) ของสัมประสิทธิ์การถดถอยของตัวแปรอิสระให้มีค่าเป็นศูนย์ คำนวณผลได้อย่างรวดเร็ว และให้ตัวแบบบางเบา (Sparse Model) ที่มีประสิทธิภาพในการพยากรณ์สูง (Bertsimas et al., 2016; Trevor Hastie et al., 2017)

แม้ว่าวิธี L1 จะมีประสิทธิภาพสูงในการพยากรณ์ แต่ยังมีข้อเสียเพราะวิธี L1 มักจะเลือกตัวแปรเข้ามาในตัวแบบมากกว่าจำนวนตัวแปรอิสระที่มีความสัมพันธ์กับตัวแปรตามอย่างแท้จริง นอกจากนี้ในกรณีที่ตัวแปรอิสระมีความสัมพันธ์กันสูง (Highly Correlated Data Set) วิธี L1 มีแนวโน้มจะคัดเลือกตัวแปรเพียงตัวเดียวจากลุ่มตัวแปรที่มีความสัมพันธ์กันสูงเข้ามาในตัวแบบ จึงทำให้ตัวแปรที่ได้มีความเอนเอียง (วิฐรา พึ่งพาพงศ์, 2558; สวรรยา ภู่เงิน, 2557)

จากข้อจำกัดด้านความไม่คงเส้นคงวาของการคัดเลือกตัวประมาณเข้าสู่ตัวแบบ Zou (Zou, 2006) ได้เสนอวิธี Adaptive Least Absolute Selection and Shrinkage Operator (Adaptive

Lasso ซึ่งในงานวิจัยนี้เรียกว่า A-L1) ซึ่งต่อยอดจากวิธี L1 คือ ยังสามารถกำหนดน้ำหนักที่ไม่เป็นลบ (Non-negative Weight) ของค่าสัมประสิทธิ์การถดถอย (β_j) แต่ละตัวให้มีค่าแตกต่างกัน ซึ่งลดความไม่คงเส้นคงวาในการประมาณค่า และแก้ปัญหาความเอนเอียงของตัวประมาณได้ อย่างไรก็ตาม วิธีนี้ยังไม่สามรถขจัดปัญหา Multicollinearity ได้

ในปี 2009 Zou และ Zhang (Zou & Zhang, 2009) ได้พัฒนาวิธี Adaptive Elastic Net Regression (Adaptive Elastic Net ซึ่งในงานวิจัยนี้เรียกว่า A-L1L2) ซึ่งเป็นที่รวมกันระหว่างวิธี A-L1 ที่มีจุดเด่นด้านความไม่เอนเอียงของตัวประมาณ และวิธี Elastic Net Regression (Elastic Net ซึ่งในงานวิจัยนี้เรียกว่า L1L2) ที่ช่วยแก้ปัญหา Multicollinearity เข้าด้วยกัน ทำให้ตัวประมาณที่ได้จากวิธี A-L1L2 เป็นตัวประมาณที่ไม่เอนเอียง และขณะเดียวกันก็สามารถคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบได้โดยไม่เกิด Multicollinearity โดยทั้ง 3 วิธีที่กล่าวมาข้างต้น ทั้งวิธี L1, A-L1 และ A-L1L2 ถือเป็น Shrinkage Method ทั้งสิ้น

ส่วนการวิเคราะห์การถดถอยสำหรับข้อมูลที่มีมิติสูง อีกวิธีหนึ่งคือ Best Subset Selection ซึ่งเป็นวิธีการที่มีมาตั้งแต่ปี 1966 (Efroymson, 1966; Draper and Smith, 1966) วิธี Best Subset Selection ใช้ L_0 norm Penalty Function (ฟังก์ชันการลงโทษด้วยจำนวนพารามิเตอร์ที่ถูกถ่วงน้ำหนัก) ในการประมาณสัมประสิทธิ์การถดถอย และในปี 2016 Bertsimas และคณะนำ Mixed Integer Optimization เข้ามาช่วยในการวิเคราะห์ พบว่าวิธี Best Subset Selection ให้ Sparse Solutions ที่มีประสิทธิภาพในการทำนายสูงกว่าวิธี L1 อย่างไรก็ตาม วิธี Best Subset Selection ยังมีข้อจำกัดเพราะสามารถใช้ได้ดีกับข้อมูลที่มีอัตราส่วนสัญญาณต่อสัญญาณรบกวน (Signal to Noise Ratio: SNR) สูง และข้อมูลจะต้องมีขนาดตัวอย่างที่พอเพียง ถึงแม้ว่าวิธีการนี้จะมีประสิทธิภาพสูงและได้ตัวแบบที่แน่นอน แต่วิธีการนี้ก็มาพร้อมกับต้นทุนการคำนวณ (Computational Cost) ที่เพิ่มขึ้นอย่างมาก ซึ่งเป็นข้อจำกัดของวิธีการ Best Subset Selection เมื่อเปรียบเทียบกับวิธีการอื่น ๆ (Hastie et al., 2017; Hazimeh & Mazumder, 2020) จนในปี 2020 Hazimeh และ Mazumder มีการพัฒนาวิธี Fast Best Subset Selection โดยใช้วิธี Coordinate Descent และ Local Combinatorial Optimization ทำให้สามารถคำนวณได้รวดเร็วขึ้น ซึ่งสามารถวิเคราะห์การถดถอยเชิงเส้นโดยใช้ L_0norm Penalty Function เพียงอย่างเดียว (ซึ่งในงานวิจัยนี้เรียกว่า L0Learn) นอกจากนี้ ยังสามารถวิเคราะห์ L_0norm Penalty Function ร่วมกับ L_1norm Penalty Function (ซึ่งในงานวิจัยนี้เรียกว่า L0L1Learn) และ L_2norm Penalty Function (ซึ่งในงานวิจัยนี้เรียกว่า L0L2Learn) ได้อีกด้วย (Hazimeh & Mazumder, 2020; Mazumder, 2020)

อย่างไรก็ตามในการศึกษาที่ผ่านมา แม้จะมีการทดลอง และเปรียบเทียบประสิทธิภาพของการวิเคราะห์การถดถอยด้วยวิธี L1, A-L1, A-L1L2, L0Learn, L0L1Learn, และ L0L2Learn อยู่

หลายครั้ง แต่ส่วนใหญ่จะให้ความสำคัญกับลักษณะของ SNR เท่านั้น แม้จะมีการสร้างข้อมูลที่ตัวแปรอิสระมีความสัมพันธ์กันในระดับต่าง ๆ ในการทดสอบ แต่ก็ยังไม่ได้มีการอภิปรายและสรุปผลอย่างชัดเจนว่าหากตัวแปรอิสระมีความสัมพันธ์กันในระดับต่างกัน แต่ละวิธีมีพฤติกรรมแตกต่างกันอย่างไร ดังนั้นในการศึกษาครั้งนี้จะสร้างข้อมูลที่มีระดับความสัมพันธ์ของตัวแปรอิสระในระดับต่าง ๆ เพื่อเปรียบเทียบประสิทธิภาพ และความถูกต้องแม่นยำของการประมาณค่าสัมประสิทธิ์การถดถอยเชิงเส้นสำหรับข้อมูลที่มีมิติสูงด้วยวิธี L1, A-L1, A-L1L2, L0Learn, และ L0L2Learn โดยศึกษาพฤติกรรมโดยละเอียดของแต่ละวิธี เพื่อให้ได้ข้อสรุปที่ชัดเจนและข้อเสนอแนะในการใช้งาน นอกจากนี้งานวิจัยที่ผ่านมาส่วนใหญ่จะเป็นการศึกษาโดยผู้คิดและพัฒนาเครื่องมือเอง การศึกษาครั้งนี้จึงเป็นการศึกษาโดยบุคคลที่สาม (Third Party) ซึ่งจะเป็นการเปรียบเทียบที่เป็นกลางสามารถอ้างอิงได้

1.2 วัตถุประสงค์การวิจัย

1) เพื่อศึกษาวิธีการประมาณค่าสัมประสิทธิ์การถดถอยเชิงเส้นสำหรับข้อมูลที่มีมิติสูงด้วยวิธี L0Learn และวิธี L0L2Learn เป็นหลัก ร่วมกับการศึกษาวิธี L1 วิธี A-L1 และวิธี A-L1L2 ซึ่งเป็นวิธีที่ได้รับความนิยมในปัจจุบัน

2) เพื่อเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าสัมประสิทธิ์การถดถอยเชิงเส้นสำหรับข้อมูลที่มีมิติสูงด้วยด้วยวิธี L0Learn และวิธี L0L2Learn ซึ่งเป็นวิธีที่ได้รับการพัฒนาใหม่ในปี 2021 โดยเปรียบเทียบกับวิธี L1 วิธี A-L1และวิธี A-L1L2 ซึ่งเป็นวิธีการประมาณค่าสัมประสิทธิ์การถดถอยเชิงเส้นสำหรับข้อมูลที่มีมิติสูงซึ่งได้รับความนิยมอย่างแพร่หลายในปัจจุบัน โดยพิจารณาประสิทธิภาพจากความสามารถในการพยากรณ์ และถูกต้องในการคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบเมื่อข้อมูลมีระดับความสัมพันธ์ของตัวแปรอิสระในระดับต่าง ๆ และมีอัตราส่วนสัญญาณต่อสัญญาณรบกวน แตกต่างกัน

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

1.4 ขอบเขตของงานวิจัย

ในการวิจัยครั้งนี้ใช้ข้อมูลที่ได้จากการจำลองในรูปแบบการความสัมพันธ์การถดถอยเชิงเส้น ซึ่งมีข้อกำหนดคือ

1) ขนาดตัวอย่าง (n) เท่ากับ 100 ตัวอย่าง
2) จำนวนตัวแปรอิสระ (p) เท่ากับ 1,000 ตัว
3) ตัวแปรอิสระมีความสัมพันธ์กันแบบยกกำลัง (Exponential Correlation) ทั้งหมด 3 ระดับ คือ

- ข้อมูลที่มีความสัมพันธ์ของตัวแปรอิสระ เท่ากับ 0.0
- ข้อมูลที่มีความสัมพันธ์ของตัวแปรอิสระ เท่ากับ 0.5
- ข้อมูลที่มีความสัมพันธ์ของตัวแปรอิสระ เท่ากับ 0.9

4) ตัวแปรอิสระ 20 ตัวแปรที่มีสัมประสิทธิ์การถดถอยไม่เท่ากับศูนย์ ประกอบไปด้วย $\beta_{1,}\beta_{2,},\beta_{3,},...,\beta_{10}=2$ และ $\beta_{101,}\beta_{102,},\beta_{103,},...,\beta_{110}=1$

5) กำหนด SNR ทั้งหมด 5 ระดับ คือ SNR = 0.1, 0.5, 1, 5, 10 และ 20

6) ค่าความแปรปรวนของความคลาดเคลื่อน (σ^2) ให้มีค่าขึ้นอยู่กับค่า SNR

7) จำลองข้อมูลจะทำซ้ำ 100 รอบในแต่ละสถานการณ์

การประมาณค่าสัมประสิทธิ์การถดถอยเชิงเส้นจะใช้การวิเคราะห์ทั้งหมด 5 วิธี ได้แก่ 1) L1 2) A-L1 3) A-L1L2 4) L0Learn และ 5) วิธี L0L2Learn ซึ่งในการวิเคราะห์การถอดถอยด้วยวิธี L1, A-L1 และ A-L1L2 จะอาศัยการวิเคราะห์โดยใช้แพ็คเกจ glmnet (Friedman et al, 2010) เวอร์ชัน 4.1-1 เป็นเวอร์ชันล่าสุดที่เผยแพร่ออกมาในปี 2021 ส่วนการวิเคราะห์การถอดถอยด้วยวิธี L0Learn และ วิธี L0L2Learn จะใช้แพ็คเกจ L0Learn เวอร์ชัน 2.0.0 เผยแพร่ในปี 2020 พัฒนาโดย Hazimeh และคณะ ซึ่งทั้ง 2 แพ็คเกจเป็นแพ็คเกจที่อยู่ในโปรแกรม R และใช้ 5-fold Cross validations ในการหาค่าพารามิเตอร์ปรับแต่งที่เหมาะสมในการวิเคราะห์ทั้ง 5 วิธี

1.5 เกณฑ์การวัดประสิทธิภาพที่ใช้ในการวิจัย

เกณฑ์ที่ใช้ในการพิจารณาเปรียบเทียบประสิทธิภาพของการประมาณค่าสัมประสิทธิ์การถดถอยด้วยวิธี L0Learn, L0L2Learn, L1, A-L1, และวิธี A-L1L2 จะพิจารณาใน 2 มิติหลัก คือ 1) ประสิทธิภาพของการพยากรณ์ และ 2) ประสิทธิภาพในการเลือกตัวประมาณ ซึ่งในแต่ละมิติมีรายละเอียดโดยย่อ ดังนี้

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

1. ประสิทธิภาพด้านการพยากรณ์

การวัดประสิทธิภาพของการพยากรณ์ ใช้บ่งบอกว่าค่าพยากรณ์ที่ได้จากการวิเคราะห์ใกล้เคียงกับค่าสังเกตมากเพียงใด วัดจากค่าเฉลี่ยความคลาดเคลื่อนกำลังสอง (Average of Mean Squared Error: AMSE) ที่ได้จากการจำลองข้อมูลทั้งหมด 100 ครั้ง

2. ประสิทธิภาพด้านเลือกตัวประมาณ

นอกจากประสิทธิภาพของการพยากรณ์แล้ว อีกประเด็นหนึ่งที่ต้องให้ความสำคัญในการวิเคราะห์ตัวแบบบางเบา (Sparse Model) คือความสามารถของตัวแบบในการคัดเลือกตัวแปรอิสระที่ถูกต้องเข้าสู่ตัวแบบ

สำหรับการวัดประสิทธิภาพในการเลือกตัวประมาณจะวัดใน 3 มิติด้วยกัน คือ

1) ค่าเฉลี่ยของค่าความแม่นยำ (Mean of Precision) เป็นการวัดความแม่นยำในการคัดเลือกตัวแปร คือ โดยเฉลี่ยแล้วตัวแบบสามารถคัดเลือกตัวแปรที่มีค่าสัมประสิทธิ์ไม่เท่ากับศูนย์ในชุดข้อมูลจริง (True Non-zero Coefficient) เป็นสัดส่วนเท่าใดเมื่อเปรียบเทียบกับจำนวนตัวแปรที่มีค่าสัมประสิทธิ์ไม่เท่ากับศูนย์ในตัวแบบ (Non-zero Coefficient)

2) ค่าเฉลี่ยของค่าความระลึก (Mean of Recall หรือ Mean of Sensitivity) เป็นการวัดความสามารถในการเรียกคืนตัวแปร คือ โดยเฉลี่ยแล้วตัวแบบสามารถเรียกคืนตัวแปรที่เป็น True Non-zero Coefficient คิดเป็นสัดส่วนเท่าใดเมื่อเทียบกับจำนวน True Non-zero Coefficient ในชุดข้อมูลจริง ซึ่งในที่นี้มี True Non-zero Coefficient ทั้งหมด 20 ตัว

3) ค่าเฉลี่ยของค่าพื้นที่ใต้กราฟ ROC (Mean of Receiver Operating Characteristic) หรือค่าเฉลี่ยของค่า AUC (Area Under the Curve) ใช้ระบุความน่าจะเป็นว่าตัวแบบสามารถแยกตัวแปรที่เป็น True Non-zero Coefficient และ True Zero Coefficient ออกจากกันได้ดีเพียงใด

1.6 ประโยชน์ที่คาดว่าจะได้รับ

1) สามารถเลือกใช้วิธีการประมาณค่าสัมประสิทธิ์การถดถอยเชิงเส้นสำหรับข้อมูลที่มีมิติสูงได้อย่างเหมาะสมกับข้อมูลที่มีความสัมพันธ์ของตัวแปรอิสระ และมีอัตราส่วนสัญญาณต่อสัญญาณรบกวนในระดับแตกต่างกัน

2) สามารถใช้อ้างอิง และเป็นแนวทางในการศึกษาวิธีการประมาณค่าสัมประสิทธิ์การถดถอยเชิงเส้นสำหรับข้อมูลที่มีมิติสูงในการเปรียบเทียบกับการประมาณด้วยวิธีอื่น ๆ ต่อไป

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

1.7 คำจำกัดความที่ใช้ในงานวิจัย

1) ข้อมูลที่มีมิติสูง (High Dimensional Data) หมายถึง ข้อมูลที่มีจำนวนตัวแปรอิสระมากกว่าขนาดตัวอย่าง

2) ฟังก์ชันลงโทษ (Penalty Function) หมายถึง ฟังก์ชันเงื่อนไขที่ใช้ในการแก้ปัญหาการหาค่าสูงสุด (Maximum Problem) หรือปัญหาการหาค่าต่ำสุด (Minimum Problem)

3) การถดถอยเชิงเส้นแบบพีนอลไลซ์ (Penalized Regression) หมายถึง วิธีการวิเคราะห์การถดถอย ภายใต้ข้อจำกัดของฟังก์ชันลงโทษ (Penalty Function)

4) พารามิเตอร์ปรับแต่ง (Tuning Parameter) หมายถึง พารามิเตอร์ที่ใช้ในการปรับลดน้ำหนัก (Shrinkage) ค่าประมาณของค่าสัมประสิทธิ์การถดถอย

5) L0Learn หมายถึง วิธีวิเคราะห์การการถดถอยแบบ Best Subset Selection ที่วิเคราะห์ภายใต้ L0Learn Algorithms

6) L0L2Learn หมายถึง วิธีวิเคราะห์การการถดถอยแบบ L_2- regularized Best Subset Selection ที่วิเคราะห์ภายใต้ L0Learn Algorithms

7) L1 หรือ Lasso หมายถึง การถดถอยเชิงเส้นพหุคูณด้วยวิธีลาซโซ (Least Absolute Selection and Shrinkage Operator: Lasso)

8) A-L1 หรือ Adaptive Lasso หมายถึง การถดถอยเชิงเส้นพหุคูณด้วยวิธีลาสโซปรับได้ (Adaptive Least Absolute Selection and Shrinkage Operator)

9) A-L1L2 หรือ Adaptive Elastic Net หมายถึง การถดถอยเชิงเส้นพหุคูณด้วยวิธีอิลาสติคเน็ตแบบปรับได้ (Adaptive Elastic Net Regression Method)

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

บทที่ 2
ทฤษฎี แนวคิดและงานวิจัยที่เกี่ยวข้อง

การวิเคราะห์การถดถอยเชิงเส้น (Linear Regression) เมื่อข้อมูลมีขนาดตัวอย่าง (n) มากกว่าจำนวนตัวแปรอิสระ (p) จะนิยมใช้วิธีกำลังสองน้อยที่สุด (Ordinary Least Squares Method: OLS) ซึ่งตัวประมาณที่ได้จะมีคุณสมบัติของตัวประมาณที่ดี คือ ไม่เอนเอียง (Unbiased) มีความแปรปรวนต่ำ (Minimum variance) มีความคงเส้นคงวา (Consistency) มีความพอเพียง (Sufficiency) และมีประสิทธิภาพสูงสุด (Efficiency) แต่เมื่อข้อมูลมีจำนวนตัวแปรอิสระ (p) มากกว่าจำนวนตัวอย่าง (n) หรือที่เรียกว่าข้อมูลที่มีมิติสูง (High Dimensional Data) จะนิยมใช้วิธี Penalized Regression เป็นวิธีการที่เพิ่มฟังก์ชันการลงโทษ (Penalty Function) เข้าไปในสมการประมาณค่าสัมประสิทธิ์การถดถอย วิธี Penalized Regression มีหลากหลายรูปแบบ แต่ในงานวิจัยฉบับนี้จะกล่าวถึงทั้งหมด 5 วิธี ได้แก่ 1) Fast Best subset selection หรือเรียกว่าวิธี L0Learn 2) L_2- regularized Best Subset Selection หรือเรียกว่าวิธี L0L2Learn 3) Least Absolute Selection and Shrinkage Operator หรือเรียกว่าวิธี L1 4) Adaptive Least Absolute Selection and Shrinkage Operator หรือเรียกว่าวิธี A-L1 และ 5) วิธี Adaptive Elastic Net Regression Method หรือเรียกว่าวิธี A-L1L2 โดยวิธีการหาค่าพารามิเตอร์ปรับแต่ง (Tuning Parameter) ของการประมาณค่าสัมประสิทธิ์การถดถอยเชิงเส้นสำหรับข้อมูลที่มีมิติสูงทั้ง 5 วิธีจะใช้วิธีการตรวจสอบไขว้ (Cross-Validation) ซึ่งในการศึกษาครั้งนี้ผู้วิจัยได้ศึกษาตำรา งานวิจัย รวมไปถึงเอกสารต่าง ๆ ซึ่งได้รวบรวมไว้ในรายละเอียดดังต่อไปนี้

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

2.1 ตัวแบบการถดถอยเชิงเส้นพหุคูณ (Multiple linear regression model)

ตัวแบบการถดถอยเชิงเส้นพหุคูณ (Multiple Linear Regression) เป็นการหาตัวแบบที่ใช้ในการอธิบายความสัมพันธ์เชิงเส้น (Liner Relationship) ระหว่างตัวแปรอิสระที่มีมากกว่า 1 ตัวแปรกับตัวแปรตามเพียง 1 ตัวผ่านตัวแบบซึ่งประกอบด้วยส่วนของฟังก์ชันที่อธิบายลักษณะความสัมพันธ์ของตัวแปรอิสระและตัวแปรตาม และส่วนที่เป็นความคลาดเคลื่อนสุ่ม ซึ่งสามารถเขียนได้ดังสมการที่ (1)

$$Y = x\beta + \varepsilon \tag{1}$$

เมื่อ $Y = (Y_1, Y_2, Y_3, ..., Y_n)^T$ เป็นเวกเตอร์ของตัวแปรตามขนาด $n \times 1$

$X = (X_1, X_2, X_3, ..., X_n)^T$ เป็นเมทริกซ์ของตัวแปรอิสระขนาด $n \times p$ โดยที่ $X_i = (X_{i1}, X_{i2}, X_{i3}, ..., X_{ip})^T$ เป็นเวกเตอร์ค่าสังเกตที่ i ของตัวแปรอิสระ p ตัว

$\beta = (\beta_1, \beta_2, \beta_3, ..., \beta_p)^T$ เป็นเวกเตอร์ค่าสัมประสิทธิ์การถดถอยขนาด $p \times 1$

$\varepsilon = (\varepsilon_1, \varepsilon_2, \varepsilon_3, ..., \varepsilon_n)^T$ เป็นเวกเตอร์ความคลาดเคลื่อนสุ่มขนาด $n \times 1$ และ $\varepsilon \sim N(0, \sigma^2 I_n)$ โดยที่แต่ละตัวมีการแจกแจงที่มีลักษณะเดียวกันและเป็นอิสระกัน (Identically Independent Distribution: i.i.d.)

จากตัวแบบการถดถอยเชิงเส้นพหุคูณในสมการที่ (1) ซึ่งเป็นตัวแบบเชิงเส้น และเทอม $X\beta$ เป็นค่าคงที่ ภายใต้สมมติฐานของ ε เป็นค่าความคลาดเคลื่อนแบบสุ่ม จะได้เวกเตอร์ Y ที่มีการแจกลักษณะเดียวกันกับ ε และเป็นอิสระ ซึ่งมีเวกเตอร์ค่าเฉลี่ย ดังสมการที่ (2)

$$\begin{aligned} E[Y] &= E[X\beta + \varepsilon] \\ E[Y] &= E[X\beta] + E[\varepsilon] \\ E[Y] &= X\beta \end{aligned} \tag{2}$$

และมีความแปรปรวนร่วม ดังสมการที่ (3)

$$\begin{aligned} \text{cov}[Y] &= \text{cov}[X\beta + \varepsilon] \\ \text{cov}[Y] &= \text{cov}[\varepsilon] \\ \text{cov}[Y] &= \sigma^2 I_n \end{aligned} \tag{3}$$

การสมมติฐานของ ε ที่ทำให้ทราบรูปแบบการแจกแจงและความแปรปรวนของตัวแปรตามแล้ว ยังทำให้สามารถหาค่าเฉลี่ยของเวกเตอร์ของตัวแปรตามเมื่อกำหนดค่า X_i ได้ ดังสมการที่ (4)

$$E[Y \mid X_i] = X_i\beta \tag{4}$$

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

2.2 การวิเคราะห์การถดถอยเชิงเส้นด้วยวิธีกำลังสองน้อยที่สุด (Least Squares Method)

วิธีการประมาณค่าพารามิเตอร์ในการวิเคราะห์การถดถอยเชิงเส้นพหุคูณมีหลายวิธี แต่วิธีที่เป็นที่นิยมใช้กันอย่างแพร่หลายคือ วิธีกำลังสองน้อยที่สุด (Least Squares Method) โดยมีหลักในการประมาณค่า คือหาตัวประมาณของเวกเตอร์ β ที่ทำให้เส้นถดถอยของตัวอย่างใกล้เคียงกับค่าสังเกตมากที่สุด โดยพิจารณาจากค่าความคลาดเคลื่อนสุ่ม ε_i ซึ่งเป็นผลต่างระหว่างค่าสังเกตที่ i ของตัวแปรตามและค่าคาดหวังที่ได้จากการประมาณของตัวแบบ สามารถเขียนอยู่ในรูปสมการได้ดังสมการที่ (5)

$$\varepsilon_i = Y_i - E[Y \mid X_i] \tag{5}$$

$$\varepsilon_i = Y_i - X_i\beta$$
$$\varepsilon_i = Y_i - \hat{Y}_i$$

ตัวประมาณที่ได้จากวิธีกำลังสองน้อยที่สุดคือตัวประมาณที่ทำให้ผลบวกกำลังสองของค่าความคลาดเคลื่อนสุ่ม ε_i (Residual Sum of Square: RSS) มีค่าต่ำที่สุด เขียนได้ดังสมการที่ (6)

$$RSS(\beta) = \sum_{i=1}^{n} (Y_i - X_i^T \beta)^2 \tag{6}$$

และสามารถเขียนให้อยู่ในรูปปัญหาการหาค่าเหมาะสมที่สุด (Optimization Problem) ดังสมการ (7)

$$\underset{\beta}{\text{minimize}} \frac{1}{2} \left\| Y - \sum_{j=j}^{p} X_j \beta_j \right\|^2 \tag{7}$$

แต่เนื่องจากฟังก์ชันผลบวกกำลังสองของค่าความคลาดเคลื่อนสุ่มเป็นฟังก์ชันกำลังสอง (Quadratic Function) ที่สามารถหาค่าต่ำสุดได้เสมอแต่ค่าที่ได้มีโอกาสเป็นได้หลายค่า (Un-unique) ซึ่งคำตอบหรือค่าประมาณพารามิเตอร์ของวิธีกำลังสองน้อยที่สุดสามารถเขียนได้ดังสมการที่ (8)

$$RSS(\beta) = (Y - X\beta)^T (Y - X\beta) \tag{8}$$

จากนั้นหาอนุพันธ์อันดับหนึ่ง (Differentiation) ของสมการที่ (8) เทียบกับ β แล้วให้เท่ากับศูนย์ จะได้สมการปกติ ดังสมการที่ (9)

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

$$X^T (Y - X\beta) = 0 \tag{9}$$
$$X^T X\beta = X^T Y$$

ถ้า $X^T X$ ไม่เป็นเมทริกซ์เอกฐาน (Nonsingular Matrix) จะได้คำตอบที่เป็นเอกลักษณ์ (Unique Solution) ของสมการ (James et al., 2013) ซึ่งเขียนได้ดังสมการที่ (10)

$$\hat{\beta} = \underset{\beta}{\arg\min} \frac{1}{2} \left\| Y - \sum_{j=1}^{p} X_j \beta_j \right\|^2 = (X^T X)^{-1} X^T Y \tag{10}$$

โดยตัวประมาณที่ได้จะเป็นตัวประมาณที่มีสมบัติเชิงเส้นตรง คือ $\hat{\beta}$ มีความสัมพันธ์เชิงเส้นตรงกับตัวแปรตาม และมีสมบัติความไม่เอนเอียง (Unbiased Estimator) คือ $\hat{\beta}$ เป็นตัวประมาณที่ไม่

เอนเอียงสำหรับ β และสุดท้ายคือ $\hat{\beta}$ มีสมบัติความแปรปรวนต่ำที่สุด (Minimum Variance) (พัชราภรณ์ พรดำเนินสวัสดิ์, 2560)

Tibshirani (Tibshirani, 1996) มีข้อสังเกตถึงตัวประมาณที่ได้จากการประมาณด้วยวิธีกำลังสองน้อยที่สุดใน 2 ประเด็นคือ 1) ความถูกต้องของการพยากรณ์ (Prediction Accuracy) แม้ว่าตัวประมาณที่ได้จากวิธีกำลังสองน้อยที่สุดมักจะเป็นตัวประมาณที่ไม่เอนเอียงแต่กลับมีความแปรปรวนสูง และ 2) ความสามารถในการอธิบายตัวแบบ เนื่องจากวิธีกำลังสองน้อยที่สุดจะให้ตัวแบบขนาดใหญ่ยากต่อการอธิบาย ซึ่งในความเป็นจริงนักวิจัยต้องการตัวแบบที่มีขนาดเล็กและตัวแปรอิสระที่อยู่ในตัวแบบควรเป็นตัวแปรที่มีอิทธิพลต่อตัวแปรตามเท่านั้น ซึ่งปัญหาทั้งสองประการสามารถแก้ไขได้ด้วยการหด (Shrinkage) หรือกำหนดให้ค่าพารามิเตอร์ β มีค่าเท่ากันศูนย์

อีกทั้งการหาค่าประมาณด้วยวิธีกำลังสองน้อยที่สุดจะทำได้ก็ต่อเมื่อข้อมูลมีจำนวนตัวแปรอิสระน้อยกว่าจำนวนตัวอย่าง ($p < n$) แต่เมื่อข้อมูลมีจำนวนตัวแปรอิสระมากกว่าจำนวนตัวอย่าง ($p > n$) จะไม่สามารถประมาณค่าด้วยวิธีกำลังสองน้อยที่สุดได้ เนื่องจากเมทริกซ์จัตุรัส ($X^T X$) จะเป็นเมทริกซ์เอกฐาน (Singular Matrix) ไม่สามารถหาเมทริกซ์ผกผัน (Inverse Matrix) ได้ นอกจากนี้ค่าประมาณที่ได้ไม่เสถียรและไม่มีประสิทธิภาพ เนื่องจากปัญหาตัวแปรอิสระมีความสัมพันธ์เชิงเส้นกันสูง (Multicollinearity) อีกทั้งตัวแบบที่ได้จะมีขนาดใหญ่ มีความซับซ้อนและอธิบายความสัมพันธ์ได้ยาก (วิฐรา พึ่งพาพงศ์, 2558)

2.2 การวิเคราะห์การถดถอยเชิงเส้นด้วยฟังก์ชันลงโทษ (Penalized Regression)

Penalized Regression หรือ Regularized Regression เป็นที่นิยมใช้อย่างแพร่หลายในการวิเคราะห์ข้อมูลมิติสูง ถูกพัฒนามาจากวิธี OLS แตกต่างกันเพียง Regularized Regression มีฟังก์ชันเงื่อนไขในการหาค่าประมาณ เรียกว่า Penalized Function (วิฐรา พึ่งพาพงศ์, 2558) โดย Regularized Regression จะแบ่งออกเป็น 2 กลุ่มใหญ่ คือ 1) Best Subset Selection ซึ่งใช้ $L_0 norm$ เป็น Penalty Function ในการหาคำตอบ คำตอบที่ได้จะเป็นค่าแน่นอน แต่จะใช้ต้นทุนในการคำนวณสูง และใช้เวลาในการคำนวณค่อนข้างนานเมื่อเปรียบเทียบกับ Shrinkage Method (James et al., 2013) ซึ่งในระยะต่อมาวิธี Best Subset Selection ก็ได้มีการปรับปรุงโดยการเพิ่ม Shrinkage Term เข้ามาในตัวแบบทำให้มีความรวดเร็วในการคำนวณมากขึ้นรวมถึงยังคงคุณสมบัติที่ดีของตัวประมาณที่ได้จากวิธี Best Subset Selection ไว้ได้ (Mazumder et al., 2017) และ 2) Shrinkage Method ซึ่งจะใช้ $L_1 norm$ และ $L_2 norm$ เป็น Penalty Function ในการหาคำตอบ ซึ่งคำตอบที่ได้จะเป็นค่าประมาณ มีจุดเด่นในด้านความรวดเร็วในการวิเคราะห์ และเป็นวิธีที่นิยมใช้กันอย่างแพร่หลาย ซึ่งวิธีที่กล่าวมาทั้งสองกลุ่มสามารถประมาณค่าสัมประสิทธิ์การถดถอยและ

ขณะเดียวกันก็สามารถคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบโดยทำให้ค่าสัมประสิทธิ์การถดถอยบางตัวมีค่าเท่ากับศูนย์ หรือที่เรียกว่าตัวแบบบางเบา (Sparse Model) ซึ่งเป็นวิธีการที่มีประสิทธิภาพในวิเคราะห์ข้อมูลที่มีมิติสูง เพราะตัวแบบที่ได้จะมีขนาดกะทัดรัด ง่ายต่อการอธิบาย (Bühlmann & Van De Geer, 2011)

2.2.1 การถดถอยเชิงเส้นพหุคูณด้วยวิธีการคัดเลือกชุดข้อมูลย่อยที่ดีที่สุด (Best Subsets Selection)

Best Subset Selection (L0) เป็น Traditional Method สำหรับการวิเคราะห์ข้อมูลมิติสูง โดยใช้ L_0norm เป็น Penalty Function ซึ่งจะได้ตัวประมาณที่มีคุณสมบัติที่ดีทางสถิติ ไม่เอนเอียง และมีความคงเส้นคงวา และได้ตัวแบบที่ถูกต้องแม่นยำ (Zhang & Zhang, 2012) ดังนั้นวิธี Best Subset Selection จึงได้รับการพัฒนามาอย่างต่อเนื่อง เพื่อให้ง่ายต่อการทำความเข้าใจ จะแบ่งการพัฒนาวิธี Best Subset Selection ออกเป็น 3 ช่วงคือ

ช่วงที่ 1 เป็น Traditional Best Subset Selection ที่ใช้หลักการสร้างตัวแบบจากทุก ๆ รูปแบบ (Combination) ของตัวแปรอิสระทั้งหมด p ตัว ซึ่งจุดอ่อนของวิธีนี้คือเป็นปัญหาแบบ NP Hard ต้องใช้ต้นทุนในการคำนวณสูง ในอดีตโปรแกรม leaps (R Package) สามารถใช้กับข้อมูลที่มีตัวแปรอิสระไม่เกิน 30 ตัว (Bertsimas et al., 2016; Mazumder et al., 2017)

ช่วงที่ 2 Modern Best Subsets Selection ช่วงประมาณปี 2016-2019 เนื่องมาจากการพัฒนาของเทคโนโลยีคอมพิวเตอร์ที่เป็นไปอย่างรวดเร็ว โดยพบว่าในปี 2015 คอมพิวเตอร์มีความเร็วในการคำนวณถึง $10^{5.57}$ เท่าเมื่อเทียบกับปี 1994 Bertsimas และคณะ (Bertsimas et al., 2016; Mazumder et al., 2017) ได้ใช้ Mixed Integer Optimization (MIO) แม้ว่าจะสามารถคำนวณได้รวดเร็วมากขึ้น และคำตอบที่ได้เป็นค่าแน่นอน (Exact Solution) แต่ข้อเสียคือ ยังไม่สามารถใช้ได้กับข้อมูลที่มีจำนวนตัวแปรอิสระมากกว่า 1000 ตัว และใช้เวลาในการคำนวณมาก (Hastie et al., 2017; Hazimeh & Mazumder, 2020; Mazumder et al., 2017)

ช่วงที่ 3 Fast Best Subsets Selection เป็นการพัฒนาต่อยอดมาจากช่วงที่ 2 และมุ่งเน้นการสร้างเครื่องมือที่สามารถคำนวณได้รวดเร็วมีต้นทุนในการคำนวณต่ำเหมือน Shrinkage Method โดยใช้ Cyclic Coordinate Descent (Cyclic CD) และ Local Combinatorial Optimization ในการคำนวณ ทำให้คงคุณสมบัติการคัดเลือกตัวแปรที่มีประสิทธิภาพของวิธี Best Subset Selection และสามารถคำนวณได้รวดเร็วแม้จะมีตัวแปรอิสระเป็นจำนวนมาก อย่างไรก็ตามคำตอบที่ได้จะเป็นค่าประมาณ (Approximate Solution) (Hazimeh & Mazumder, 2020; Hazimeh et al., 2022)

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

โดยแต่ละช่วงของการพัฒนาจะมีรายละเอียด ดังต่อไปนี้

1) Traditional Best Subsets Selection

วิธี Best Subsets Selection หรือ L_0- Regularized Regression เป็นการหาชุดตัวแปรย่อยที่ดีที่สุดโดยใช้ฟังก์ชันการลงโทษด้วยจำนวนพารามิเตอร์ที่ถูกถ่วงน้ำหนักหรือที่เรียกว่า L_0norm Penalty Function ซึ่งสามารถเขียนให้อยู่ในรูปสมการได้ ดังสมการที่ (11)

$$\min_{\beta}\left\{\frac{1}{2}\left\|Y-\sum_{j=j}^{p}X_j\beta_j\right\|^2+\lambda_0\|\beta\|_0\right\} \tag{11}$$

เมื่อ $\|\beta\|_0$ คือ จำนวนตัวแปรที่มีค่าสัมประสิทธิ์ไม่เท่ากับศูนย์ในตัวแบบ

λ_0 คือ พารามิเตอร์ปรับแต่งที่ใช้ในการควบคุมจำนวนตัวแปรที่มีค่าสัมประสิทธิ์ไม่เท่ากับศูนย์ในตัวแบบ โดยที่ $\lambda_0 \geq 0$

หลักการของวิธี Best Subsets Selection คือการหาค่าของตัวประมาณด้วยวิธีกำลังสองน้อยที่สุดในทุก ๆ รูปแบบ (Combination) ของตัวแปรอิสระทั้งหมด p ตัว คือ สร้างตัวแบบที่ไม่มีตัวแปรอิสระในตัวแบบเลย ต่อมาสร้างตัวแบบที่มีตัวแปรอิสระเพียง 1 ตัวจะได้ทั้งหมด p ตัวแบบ จากนั้นสร้างตัวแบบที่มีตัวแปรอิสระ 2 ตัว ซึ่งจะได้ตัวแบบทั้งหมด $\binom{p}{2}=\frac{p(p-1)}{2}$ ตัวแบบ จากนั้นสร้างตัวแบบที่มีจำนวนตัวแปรอิสระมากขึ้นจนกระทั่งมีตัวแปรอิสระทั้งหมด p ตัวในตัวแบบซึ่งในกรณีนี้จะได้ตัวแบบเพียง 1 ตัวแบบเท่านั้น จากนั้นคัดเลือกตัวแปรที่มีค่า RSS ต่ำที่สุดในแต่ครั้งของการเพิ่มตัวแปรอิสระ และในขั้นตอนสุดท้ายใช้การทำ Cross Validation ในเลือกตัวแบบที่ดีที่สุดจากตัวแบบทั้งหมด p รูปแบบ (James et al., 2013)

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

Algorithms of Best Subset Selection

1. กำหนดให้ M_0 คือตัวแบบที่ไม่ตัวแปรอิสระ (Null Model) ซึ่งเป็นตัวแบบอย่างง่ายที่สุด
2. เมื่อ $k=1,2,3,...,p$
   - สร้างตัวแบบ $\binom{p}{k}$ คือ เมื่อมีตัวแปรอิสระทั้งหมด p จะมีตัวแปรอิสระจำนวน k ตัวในตัวแบบ
   - การสร้างตัวแบบ $\binom{p}{k}$ ในแต่ละ k จะเลือกตัวแบบที่มีค่า RSS ต่ำที่สุดและแทนตัวแบบด้วย M_k
3. เลือกตัวแบบที่ดีที่สุดจากตัวแบบ $M_0,M_1,M_2,...,M_p$ โดยเลือกจากค่า AIC, BIC หรือค่า R-square ที่ได้จากการทำ Cross Validation

การประมาณค่าของตัวประมาณด้วยวิธี Best Subset Selection แม้จะได้ตัวประมาณที่มีคุณสมบัติที่ดีทางสถิติ ไม่เอนเอียงและมีความคงเส้นคงวา ตัวแบบมีความถูกต้องแม่นยำ (Raskutti et al., 2011; Zhang & Zhang, 2012) แต่ยังมีข้อจำกัดหลายประการ คือ ในการเลือกตัวแบบที่ดีที่สุดเราเลือกจากค่า RSS ที่ต่ำที่สุด หรือค่า R-square ที่สูงที่สุด แต่โดยทั่วไปตัวแบบที่มีจำนวนตัวแปรอิสระมากมักจะมีค่า RSS ต่ำหรือค่า R-square สูง (Hastie et al., 2009; James et al., 2013) ดังนั้นตัวแบบสุดท้ายที่ได้มักจะเป็นตัวแบบที่มีตัวแปรจำนวนมากที่สุดเสมอ ซึ่งตัวแบบลักษณะนี้จะมีข้อจำกัดด้านการอธิบายและตัวแปรอิสระในตัวแบบมีความสัมพันธ์กันเอง (Multicollinearity)

นอกจากนี้ค่า RSS ในข้อมูลชุดทดลอง (Training Data) จะมีต่ำ แต่เมื่อทดสอบตัวแบบกับข้อมูลชุดทดสอบ (Test Data) จะพบว่าเกิดความคลาดเคลื่อนในการพยากรณ์ค่อนข้างสูง หรือที่เรียกว่า Overfitting ซึ่งเป็นผลลัพธ์ที่ไม่พึงประสงค์ในการสร้างตัวแบบ (James et al., 2013; Tibshirani, 1996) อีกทั้งการหาค่าของตัวประมาณด้วยวิธี Best Subset Selection เป็นปัญหาแบบ NP Hard ต้องใช้ต้นทุนในการคำนวณสูง ในอดีตโปรแกรม leaps (R Package) สามารถใช้กับข้อมูลที่มีตัวแปรอิสระไม่เกิน 30 ตัว (Bertsimas et al., 2016; Hastie et al., 2017) และมักจะเกิด Overfitting โดนเฉพาะข้อมูลที่มี SNR ต่ำ (Hastie et al., 2017; Mazumder et al., 2017)

2) Modern Best Subsets Selection

จากปัญหาต้นทุนในการคำนวณสูงของวิธี Best Subset Selection ในปี 2016 Bertsimas และคณะ (Bertsimas et al., 2016) ได้พัฒนาวิธีแก้ปัญหา Best Subsets Selection ที่มีประสิทธิภาพและมีความเร็วในการคำนวณมากขึ้นโดยการนำ Mixed Integer Program เข้ามาช่วยในการแก้ปัญหา โดยการเปลี่ยนรูปปัญหาให้อยู่ในรูปปัญหาการหาค่าเหมาะสมที่สุดแบบผสม (Mixed Integer Optimization Problem) ซึ่งเขียนได้ดังสมการ (12)

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

$$\min_{\beta} \frac{1}{2}\left\| Y - \sum_{j=j}^{p} X_j \beta_j \right\|^2 \tag{12}$$

$$\text{subject to} \quad -MZ_i \le \beta_i \le MZ_i, i = 1,2,3,...,p$$

$$Z_i \in \{0,1\}, i = 1,2,3,...,p$$

$$\sum_{i=1}^{p} Z_i \le k$$

เมื่อ Z เป็นตัวแปรไบนารี (Binary Variable)

M คือ ค่าคงที่

แม้ว่าจะสามารถคำนวณได้รวดเร็วมากขึ้น โดยสามารถคำนวณได้ปัญหาที่มีจำนวนตัวแปรอิสระประมาณ 1000 ตัวแปร $(p \sim 1000)$ ในเวลาไม่กี่นาที แต่เมื่อเทียบกับวิธีอื่น ๆ เช่น L1 กลับพบว่า วิธี L1 สามารถคำนวณปัญหาที่มีขนาดใหญ่กว่าได้ภายในไม่กี่วินาที ดังนั้นวิธี MIO จึงเหมาะกับปัญหาขนาดเล็กถึงขนาดกลางเท่านั้น (Hazimeh & Mazumder, 2020)

ต่อมาในปี 2017 Mazumder และคณะ (Mazumder et al., 2017) ได้เสนอแนวทางในการแก้ปัญหา Overfitting ของวิธี Best Subset Selection เมื่อ SNR มีค่าต่ำนั้น โดยวิธีแก้ปัญหาคือ นอกเหนือจากการใช้ฟังก์ชันลงโทษ $L_0 norm$ เพื่อกำหนดจำนวนตัวแปรที่มีค่าสัมประสิทธิ์ไม่เท่ากับศูนย์ (Non-zero Coefficient) เพียงอย่างเดียว ในการวิเคราะห์ได้เพิ่ม $L_1 norm$ หรือ $L_2 norm$ เข้ามาในตัวแบบเพื่อกำหนดขนาดของการการหด (Shrink) ของค่าประมาณ ซึ่งสามารถลดการเกิด Overfitting ได้ ซึ่งสามารถเขียนปัญหาในรูปสมการได้ดังสมการที่ (13)

$$\min_{\beta} \frac{1}{2}\left\| Y - \sum_{j=j}^{p} X_j \beta_j \right\|^2 + \lambda_q \|\beta\|_q^q \tag{13}$$

$$subject\ to\ \|\beta\|_0 \le k$$

เมื่อ $q = 0,1$

$\lambda_q \ge 0$ เป็นพารามิเตอร์การปรับแต่ง (Tuning Parameter)

และสามารถเขียนอยู่ในรูปสมการที่สอดคล้องกัน ได้ดังสมการที่ (14)

$$\min_{\beta} \frac{1}{2}\left\| Y - \sum_{j=j}^{p} X_j \beta_j \right\|^2 + \lambda_q \|\beta\|_q^q + \lambda_0 \|\beta\|_0 \tag{14}$$

Chulalongkorn University

แต่ปัญหาในสมการที่ (13) ซึ่งเขียนในรูปสมการที่สอดคล้องกันดังสมการที่ (14) เป็นปัญหา Nonconvex Optimization เพื่อให้ง่ายต่อการประมาณค่า เพิ่มประสิทธิภาพและลดต้นทุนในการคำนวณลงจึงแปลงให้อยู่ในรูปของ Mixed Integer Optimization (MIO) ซึ่งสามารถเขียนได้ดังสมการที่ (15)

$$\min_{\beta} \frac{1}{2}\left\| Y - \sum_{j=j}^{p} X_j \beta_j \right\|^2 + \lambda_q \|\beta\|_q^q \tag{15}$$

$$subject\ to\ \ -MZ_i \le \beta_i \le MZ_i, i = 1,2,3,...,p$$

$$Z_i \in \{0,1\}, i = 1,2,3,...,p$$

$$\sum_{i=1}^{p} Z_i \leq k$$

เมื่อ Z เป็นตัวแปรไบนารี (Binary Variable)

M คือ ค่าคงที่

โดยจะให้ β และ Z เป็นพารามิเตอร์ที่ต้องการหาค่าที่เหมาะสม (Optimization Parameter) โดยตัวแปร Z_i จะเป็นตัวกำหนดว่าตัว β_i จะมีค่าเท่ากับศูนย์หรือไม่ เมื่อ $M < \infty$ และ M มีขนาดใหญ่พอคำตอบที่ได้จากการแก้ด้วย MIO ในสมการที่ (15) จะเป็นคำตอบของสมการที่ (14) ซึ่งคำตอบสามารถเขียนได้ดังสมการที่ (16)

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \frac{1}{2} \left\| Y - \sum_{j=j}^{p} X_j \beta_j \right\|^2 + \lambda_q \|\beta\|_q^q + \lambda_0 \|\beta\|_0 \qquad (16)$$

เมื่อ $\|\beta\|_0$ คือ จำนวนตัวแปรที่มีค่าสัมประสิทธิ์ไม่เท่ากับศูนย์ในตัวแบบ

$$\|\beta\|_q^q = \sum_{j=1}^{p} \left|\beta_j\right| \text{ เมื่อ } q = 1$$

$$\|\beta\|_q^q = \sum_{j=1}^{p} \beta_j^2 \text{ เมื่อ } q = 2$$

λ_0, λ_2 คือ พารามิเตอร์ปรับแต่งที่ โดยที่ $\lambda_0 \geq 0$ และ $\lambda_2 \geq 0$

อย่างไรก็ตามการหาค่าตัวประมาณด้วยวิธี Best Subset Selection ที่ผ่านมา เป็นการหาคำตอบแบบค่าแน่นอน (Exact Algorithms) ทำให้ยังมีต้นทุนในการคำนวณสูง และใช้เวลาในการคำนวณค่อนข้างนานเมื่อเปรียบเทียบกับการหาคำตอบแบบค่าประมาณ (Approximated Algorithms)

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

3) Fast Best Subsets Selection

ในปี 2020 Hazimeh และ Mazumdera (Hazimeh & Mazumder, 2020) ได้พัฒนา **“L0Learn Algorithms”** โดยใช้แนวคิดของ Mazumder และคณะในปี 2017 (Mazumder et al., 2017) คือใช้ L_0norm ร่วมกับ Shrinkage Function ได้แก่ L_1norm และ L_2norm เพื่อลดปัญหา Overfitting ร่วมกับการประยุกต์เอาแนวคิดของการหาคำตอบแบบค่าประมาณ (Approximated Algorithms) และใช้หลักการ Combination of Coordinate Descent (CD) และ Local Combinatorial Optimization เข้ามาใช้ในการคำนวณ โดย L0Learn Algorithms

จะให้คำตอบแบบค่าประมาณที่มีประสิทธิภาพทั้งด้านการพยากรณ์และการคัดเลือกตัวแปร และในการคำนวณให้มีความรวดเร็วมากขึ้นจนสามารถเรียกได้ว่าเป็น “**Fast Best Subset Selection**” (*Hazimeh & Mazumder, 2020; Hazimeh et al., 2022*)

ในการศึกษาครั้งนี้จะใช้ L0Learn Algorithms ในการวิเคราะห์การถดถอย และใช้วิธีที่อยู่ในกลุ่มของวิธี Best Subset selection ทั้งหมด 2 รูปแบบ คือ 1) ใช้วิธี L0Learn ที่มี L_0norm เป็น Regularizes Function เพียงอย่างเดียว ซึ่งเขียนคำตอบของค่าประมาณได้ดังสมการ (17)

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \frac{1}{2}\left\| Y - \sum_{j=j}^{p} X_j \beta_j \right\|^2 + \lambda_0 \|\beta\|_0 \tag{17}$$

เมื่อ $\|\beta\|_0$ คือ จำนวนตัวแปรที่มีค่าสัมประสิทธิ์ไม่เท่ากับศูนย์ในตัวแบบ

λ_0 คือ พารามิเตอร์ปรับแต่งที่ใช้ในการควบคุมจำนวนตัวแปรที่มีค่าสัมประสิทธิ์ไม่เท่ากับศูนย์ในตัวแบบ โดยที่ $\lambda_0 \geq 0$

และ 2) L0L2Learn (L_2- Regularized Best Subset Selection) คือใช้ L_0norm เป็น Regularize Function ในการคัดเลือกตัวแปรเข้าสู่ตัวแบบควบคู่ไปกับการใช้ L_2norm เป็น Regularize Function เพื่อหด (Shrink) ค่าประมาณ ซึ่งสามารถเขียนคำตอบของค่าประมาณได้ดังสมการ (18)

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \frac{1}{2}\left\| Y - \sum_{j=j}^{p} X_j \beta_j \right\|^2 + \lambda_q \|\beta\|_q^q + \lambda_0 \|\beta\|_0 \tag{18}$$

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

เมื่อ $\|\beta\|_0$ คือ จำนวนตัวแปรที่มีค่าสัมประสิทธิ์ไม่เท่ากับศูนย์ในตัวแบบ

$\|\beta\|_2 = \sum_{j=1}^{p} \beta_j^2$

λ_0, λ_2 คือ พารามิเตอร์ปรับแต่งที่ โดยที่ $\lambda_0 \geq 0$ และ $\lambda_2 \geq 0$

2.2.2 การถดถอยเชิงเส้นพหุคูณด้วยวิธีลาสโซ (Least Absolute Selection and Shrinkage Operator)

จากข้อจำกัดด้านความถูกต้องของการพยากรณ์ (Prediction Accuracy) และความสามารถในการอธิบายตัวแบบที่ได้จากการวิเคราะห์ด้วยวิธีกำลังสองน้อยที่สุด Tibshirani (Tibshirani, 1996) และ Chen (Chen et al., 1998) จึงได้เสนอวิธี Least Absolute Selection and Shrinkage Operator หรือ Lasso (ซึ่งในงานวิจัยนี้เรียกว่า L1) ซึ่งมีหลักในการการประมาณค่าสัมประสิทธิ์การ

ถดถอยคล้ายคลึงกับการประมาณด้วยวิธีกำลังสองน้อยที่สุด แต่เพิ่มฟังก์ชันการลงโทษด้วยผลรวมขนาดพารามิเตอร์ที่ถูกถ่วงน้ำหนัก หรือ L_1*norm Penalty Function* ทำให้ค่าสัมประสิทธิ์การถดถอยของตัวแปรอิสระทุก ๆ ตัวลดลง จนทำให้บางตัวมีค่าเท่ากับศูนย์หรือเรียกว่าตัวแบบบางเบา (Sparse Model) ตัวแบบที่ได้จึงง่ายต่อการอธิบาย และได้รับความนิยมใช้ในการวิเคราะห์อย่างต่อเนื่องมาจนปัจจุบัน (Friedman et al., 2010)

การวิเคราะห์การถดถอยด้วยวิธี L1 จะหาค่าตัวประมาณของเวกเตอร์ β ที่ทำให้ผลบวกกำลังสองของความคลาดเคลื่อนมีค่าต่ำที่สุดภายใต้เงื่อนไข L_1*norm* ซึ่งเป็นฟังก์ชันที่เขียนให้อยู่ในรูปของ β และมีพารามิเตอร์ λ มีค่ามากกว่าหรือเท่ากับศูนย์ สามารถเขียนอยู่ในรูปสมการ ดังสมการที่ (19)

$$\min_{\beta} \frac{1}{2}\left\| Y - \sum_{j=j}^{p} X_j \beta_j \right\|^2 \tag{19}$$

$$\textit{subject to} \sum_{j=1}^{p} \left|\beta_j\right| \le t \text{ หรือเทียบเท่ากับฟังก์ชันลงโทษ } \lambda \sum_{j=1}^{p} \left|\beta_j\right|$$

และสามารถเขียนอยู่ในรูปสมการที่สอดคล้องกันดังสมการที่ (20)

$$\min_{\beta} \frac{1}{2}\left\| Y - \sum_{j=j}^{p} X_j \beta_j \right\|^2 + \lambda \|\beta\|_1 \tag{20}$$

เมื่อ λ คือ พารามิเตอร์การปรับแต่ง (Tuning Parameter) โดยที่ $\lambda \ge 0$ และเป็นค่าคงที่

$$\|\beta\|_1 = \sum_{j=1}^{p} \left|\beta_j\right|$$

λ เป็นพารามิเตอร์การปรับแต่ง (Tuning Parameter) หาได้จากวิธีการตรวจสอบไขว้ซึ่งจะอธิบายในหัวข้อ 2.6 โดยที่ λ ใช้ในการควบคุมขนาดการหดตัว (Shrinkage) ของค่าสัมประสิทธิ์การถดถอย $\hat{\beta}$ หากกำหนดให้ $\hat{\beta}_j^0$ เป็นค่าสัมประสิทธิ์การถดถอยที่ได้จากการประมาณด้วยวิธีกำลังสองน้อยที่สุดและให้ $\lambda_0 = \sum \left|\hat{\beta}_j^0\right|$ เมื่อ $\lambda < \lambda_0$ จะทำให้ค่าสัมประสิทธิ์การถดถอย $\hat{\beta}$ ลดลง โดยที่ค่า $\hat{\beta}$ ทุกตัวจะลดลงในขนาดที่เท่ากัน ทำให้ค่า $\hat{\beta}$ บางตัวมีค่าเท่ากับศูนย์และบางตัวมีค่าไม่เท่ากับศูนย์ และหาก $\lambda = 0$ ผลลัพธ์ที่ได้จากวิธี L1 จะเหมือนกับผลลัพธ์ที่ได้จากวิธีกำลังสองน้อยที่สุด ซึ่งผลคำตอบสามารถเขียนให้อยู่ในรูปสมการที่สอดคล้องกัน ดังสมการที่ (21)

Chulalongkorn University

$$\hat{\beta} = \arg\min_{\beta} \frac{1}{2} \left\| Y - \sum_{j=j}^{p} X_j \beta_j \right\|^2 + \lambda \|\beta\|_1 \quad (21)$$

แต่เนื่องจากตัวประมาณที่ได้จากวิธี L1 เป็นคำตอบของฟังก์ชันที่ไม่เป็นเชิงเส้น (Non-Linear Function) และไม่สามารถหาค่าอนุพันธ์เทียบกับค่า β ได้ จึงไม่สามารถเขียนตัวประมาณ Lasso ให้อยู่ในรูปทั่วไปได้ แต่สามารถหาตัวประมาณจากวิธี L1 ในกรณีเฉพาะ คือ เมทริกซ์ของตัวแปรอิสระอยู่ในรูปเมทริกซ์ตั้งฉากปกติ (Orthonormal Design Matrix)

การหาค่าสัมประสิทธิ์การถดถอยด้วยวิธี L1 สามารถประมาณค่า พารามิเตอร์และสามารถคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบได้พร้อม ๆ กัน โดยใช้กระบวนการหดค่าสัมประสิทธิ์แบบต่อเนื่อง (Continuous Shrinkage) ทำให้ความถูกต้องของการพยากรณ์เพิ่มมากขึ้น มีความรวดเร็วในการคำนวณสูง และมีซอฟท์แวร์ทางสถิติที่รอบรับการการวิเคราะห์ เช่น glmnet ทำให้เป็นที่นิยมใช้อย่างแพร่หลาย

วิธี L1 มีข้อจำกัดเนื่องจากตัวประมาณมีความเอนเอียง (Biased Estimator) ในกรณีที่ค่าสัมประสิทธิ์การถดถอยมีขนาดใหญ่ (Fan & Li, 2001) และมีความไม่คงเส้นคงวาของการคัดเลือกตัวแปรเข้าสู่ตัวแบบ แม้จะใช้พารามิเตอร์ปรับแต่งที่มีความเหมาะสม (Meinshausen & Bühlmann, 2006; Zou, 2006) วิธี L1 สามารถใช้ได้ดีในกรณีที่ข้อมูลมีอัตราส่วนสัญญาณต่อสัญญาณรบกวน (Signal to Noise Ratio: SNR) ต่ำ (Zou, 2006) นอกจากนี้ยังสามารถเลือกตัวแปรเข้าสู่ตัวแบบได้มากที่สุดเท่ากับจำนวนตัวอย่าง $(p = n)$ และหากตัวแปรอิสระมีความสัมพันธ์กันสูง วิธี L1 มีแนวโน้มที่จะเลือกตัวแปรเพียงตัวเดียวจากกลุ่มตัวแปรนั้น ๆ ดังนั้น หากข้อมูลที่นำมาวิเคราะห์มีจำนวนตัวแปรอิสระมากกว่าขนาดตัวอย่างมาก ๆ หรือตัวแประอิสระมีความสัมพันธ์กันเองสูง ตัวแบบที่ได้จากการวิเคราะห์ด้วยวิธี L1 ก็อาจไม่มีความเหมาะสมในการวิเคราะห์ข้อมูล (วิฐรา พึ่งพาพงศ์, 2558)

2.2.3 การถดถอยเชิงเส้นพหุคูณด้วยวิธีลาสโซปรับได้ (Adaptive Lasso Regression)

จากข้อจำกัดด้านความไม่คงเส้นคงวาของการคัดเลือกตัวประมาณเข้าสู่ตัวแบบของวิธี L1 ทำให้ในปี 2006 Zou (Zou, 2006) ได้พัฒนาวิธี Adaptive Least Absolute Selection and Shrinkage Operator (ซึ่งในงานวิจัยนี้เรียกว่า A-L1) ซึ่งต่อยอดจากวิธี L1 คือ นอกจากการปรับน้ำหนักของ Penalty Function ด้วยพารามิเตอร์ λ แล้วยังสามารถกำหนดน้ำหนักที่ไม่เป็นลบ (Non-negative Weight) ของ β_j แต่ละตัวให้มีค่าแตกต่างกัน โดยกำหนดค่าถ่วงน้ำหนักให้มีค่าสูง

เมื่อค่าสัมประสิทธิ์มีค่าน้อย และกำหนดค่าถ่วงน้ำหนักให้มีค่าต่ำเมื่อค่าสัมประสิทธิ์มีค่ามาก เพื่อลดความไม่คงเส้นคงวาในการประมาณค่าด้วยวิธี L1 ทำให้สามารถแก้ปัญหาความเอนเอียงของตัวประมาณและเพิ่มความแม่นยำในการคัดเลือกตัวแปร ซึ่งสมการในการหาค่าของตัวประมาณสามารถเขียนอยู่ในรูปสมการ ดังสมการที่ (22)

$$\min_{\beta} \frac{1}{2}\left\| Y - \sum_{j=j}^{p} X_j \beta_j \right\|^2 \tag{22}$$

subject to $\sum_{j=1}^{p} w_j \left|\beta_j\right| \le t$ หรือเทียบเท่ากับฟังก์ชันลงโทษ $\lambda \sum_{j=1}^{p} w_j \left|\beta_j\right|$

และสามารถเขียนอยู่ในรูปสมการที่สอดคล้องกันดังสมการที่ (23)

$$\min_{\beta} \frac{1}{2}\left\| Y - \sum_{j=j}^{p} X_j \beta_j \right\|^2 + \lambda P(\mathrm{w}_j \left\|\beta_j\right\|_1) \tag{23}$$

เมื่อ λ คือ พารามิเตอร์การปรับแต่ง (Tuning Parameter) โดยที่ $\lambda \ge 0$ และเป็นค่าคงที่

w_j คือ น้ำหนักที่ทราบค่าของ β_j โดยที่ $\mathrm{w}_j \ge 0$

$$\left\|\beta\right\|_1 = \sum_{j=1}^{p} \left|\beta_j\right|$$

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

ในการกำหนดเวกเตอร์น้ำหนัก w สามารถใช้ค่าสัมประสิทธิ์การถดถอยที่ได้จากวิธีกำลังสองน้อยที่สุด $\hat{\beta}_j^{ols}$ มาใช้ในการกำหนดเวกเตอร์น้ำหนักในกรณีที่ตัวแปรอิสระไม่มีปัญหา Multicollinearity แต่โดยทั่วไปจะนิยมค่าสัมประสิทธิ์การถดถอยที่ได้จากการประมาณวิธี Ridge Regression เนื่องจากมีความเสถียร (Stable) มากกว่า

ซึ่งการศึกษาครั้งนี้กำหนดเวกเตอร์น้ำหนัก w ด้วยค่าประมาณที่ได้จากการประมาณวิธี Ridge Regression เมื่อกำหนดให้ $\gamma > 0$ จะให้ทำให้สามารถหาค่าประมาณของค่าเวกเตอร์น้ำหนัก w ดังสมการที่ (24)

$$\hat{\mathrm{w}}_j = \frac{1}{\left|\hat{\beta}_j^{Ridge}\right|^{\gamma}} \tag{24}$$

ทำให้สามารถเขียนค่าประมาณของสัมประสิทธิ์ที่ได้ดังสมการ (25)

$$\hat{\beta}=\underset{\beta}{\arg\min}\frac{1}{2}\left\|Y-\sum_{j=j}^{p}X_j\beta_j\right\|^2+\lambda P(\mathrm{w}_j\left\|\beta_j\right\|_1) \tag{25}$$

ตัวประมาณที่ได้จากวิธี A-L1 กล่าวได้ว่าเป็นตัวประมาณที่มีคุณสมบัติในการพยากรณ์ (Oracle Properties) เพราะเป็นตัวประมาณที่มีความคงเส้นคงวาในการคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบ และตัวประมาณที่ได้มีการแจกแจงลู่เข้าสู่การแจกแจงปกติ (Asymptotic Normality) เมื่อเปรียบเทียบกับวิธีอื่น ๆ ที่ให้ตัวประมาณที่มีคุณสมบัติด้านการประมาณ เช่น SCAD พบว่าวิธี A-L1 มีความเหนือกว่าในด้านความเร็วในการคำนวณและมีต้นทุนในการคำนวณต่ำเช่นเดียวกับวิธี L1 (Huang et al., 2008; Zou, 2006) นอกจากนี้แล้วหากมีจำนวนค่าสังเกต (n) มากพอ วิธี A-L1 จะมีประสิทธิภาพในการคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบดีมากเสมือนกับทราบตัวแบบที่แท้จริง (True model) (Zou, 2006) อย่างไรก็ตามวิธี A-L1 ยังไม่สามารถแก้ปัญหาเรื่อง Multicollinearity (Zou & Zhang, 2009)

2.2.4 การถดถอยเชิงเส้นพหุคูณด้วยวิธีอิลาสติคเน็ตแบบปรับได้ (Adaptive Elastic Net Regression Method, Adaptive Elastic Net)

เนื่องจากวิธี A-L1 สามารถแก้ปัญหาด้านความเอนเอียงของตัวประมาณเท่านั้น แต่ไม่สามารถแก้ปัญหา Multicollinearity ในการคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบได้ ในปี 2005 Zou และ Hastie (Zou & Hastie, 2005) ได้เสนอวิธี Elastic Net Regression Method (Elastic Net s หรือในงานวิจัยนี้เรียกว่าวิธี L1L2) โดยมีการปรับจากวิธี L1 คือเพิ่มฟังก์ชันลงโทษด้วยผลรวมขนาดพารามิเตอร์กำลังสอง หรือที่เรียกว่า L_2-norm Penalty function ซึ่งสามารถให้อยู่ในรูปสมการได้ดัง สมการที่ (26)

$$\min_{\beta}\frac{1}{2}\left\|Y-\sum_{j=j}^{p}X_j\beta_j\right\|^2+\lambda_1\left\|\beta_j\right\|_1+\lambda_2\left\|\beta_j\right\|_2^2 \tag{26}$$

เมื่อ λ_1,λ_2 คือ พารามิเตอร์การปรับแต่ง (Tuning Parameter) โดยที่ $\lambda_1,\lambda_2\geq 0$ และเป็นค่าคงที่

$$\|\beta\|_1=\sum_{j=1}^{p}\left|\beta_j\right|$$

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

$$\|\beta\|_2^2 = \sum_{j=1}^{p} \beta_j^2$$

โดยในการวิเคราะห์จะแบ่งเป็น 2 ขั้นตอนคือ เริ่มต้นด้วยการใช้ $L_2 norm$ ในการหาค่าสัมประสิทธิ์ของตัวแปร ซึ่งขั้นตอนนี้จะช่วยขจัดปัญหา Multicollinearity หลังจากนั้นจึงใช้ $L_1 norm$ มาช่วยในเรื่องการคัดเลือกตัวแปรเข้าสู่ตัวแบบ ส่วน $L_2 norm$ ซึ่งสามารถใช้ในการวิเคราะห์ข้อมูลที่มีมิติสูง (Zou & Hastie, 2005) แม้ว่าวิธี Elastic Net จะสามารถคัดเลือกตัวแปรเข้าสู่ตัวแบบได้และไม่มีปัญหา Multicollinearity แต่ยังมีข้อจำกัดคือตัวประมาณที่ได้ไม่มีคุณสมบัติการพยากรณ์และเป็นตัวประมาณเอนเอียง (Zou & Zhang, 2009)

ในปี 2009 Zou และ Zhang (Zou & Zhang, 2009) ได้พัฒนาวิธี Adaptive Elastic Net Regression Method (Adaptive Elastic Net หรือในงานวิจัยนี้เรียกว่าวิธี A-L1L2) ซึ่งเป็นที่รวมกันระหว่างวิธี A-L1 ที่มีจุดเด่นด้านความไม่เอนเอียงของตัวประมาณ และวิธี L1L2 ที่ช่วยแก้ปัญหา Multicollinearity เข้าด้วยกัน ทำให้ตัวประมาณที่ได้จากวิธี A-L1L2 เป็นตัวประมาณที่ไม่เอนเอียง และขณะเดียวกันก็สามารถคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบได้โดยไม่เกิด Multicollinearity ซึ่งสามารถเขียนสมการหาค่าประมาณได้ดังสมการที่ (27)

$$\min_{\beta} \frac{1}{2}\left\| Y - \sum_{j=j}^{p} X_j \beta_j \right\|^2 + \lambda_1 P(\mathrm{w}_j \left\| \beta_j \right\|_1) + \lambda_2 \left\| \beta_j \right\|_2^2 \qquad (27)$$

การหาค่าของตัวประมาณด้วยวิธี A-L1L2 จะเริ่มจากหาตัวประมาณจากวิธี L1L2 ก่อน คือ ใช้ $L_2 norm$ ในการหาค่าสัมประสิทธิ์ของตัวแปรแล้วใช้ $L_1 norm$ มาช่วยในเรื่องการคัดเลือกตัวแปรเข้าสู่ตัวแบบ จากนั้นจึงใส่เวกเตอร์น้ำหนัก w โดยกำหนดจากสัมประสิทธิ์การถดถอยที่หาได้จากวิธี Ridge Regression เมื่อกำหนดให้ $\gamma > 0$ จะให้สามารถค่าเวกเตอร์น้ำหนัก w ดังสมการที่ (28)

$$\hat{\mathrm{w}}_j = \frac{1}{\left|\hat{\beta}_j^{Ridge}\right|^{\gamma}} \qquad (28)$$

ซึ่งสามารถเขียนผลคำตอบหรือค่าประมาณพารามิเตอร์ของวิธี A-L1L2 ได้ดังสมการที่ (29)

$$\hat{\beta} = \arg\min_{\beta} \frac{1}{2}\left\| Y - \sum_{j=j}^{p} X_j \beta_j \right\|^2 + \lambda_1 P(\mathrm{w}_j \left\| \beta_j \right\|_1) + \lambda_2 \left\| \beta_j \right\|_2^2 \qquad (29)$$

เมื่อ $\lambda_1,\ \lambda_2$ คือ พารามิเตอร์การปรับแต่ง (Tuning Parameter) โดยที่ $\lambda_1,\ \lambda_2 \geq 0$ และเป็นค่าคงที่

$$\|\beta\|_1 = \sum_{j=1}^{p} |\beta_j|$$

$$\|\beta\|_2 = \sum_{j=1}^{p} \beta_j^2$$

ตัวประมาณที่ได้จากวิธี A-L1L2 เป็นตัวประมาณที่ไม่เอนเอียง มีความคงเส้นคงวาในการคัดเลือกตัวแปรเข้าสู่ตัวแบบ และมีประสิทธิภาพในการพยากรณ์สูง หรือกล่าวได้ว่ามี Oracle Properties นอกจากนี้ยังสามารถขจัดปัญหา Multicollinearity ได้ และเมื่อเปรียบเทียบกับวิธี L1 และวิธี A-L1 แล้วพบว่ามีประสิทธิภาพมากกว่าอย่างมีนัยสำคัญทางสถิติ อย่างไรก็ตามวิธี A-L1L2 จะใช้ได้อย่างมีประสิทธิภาพมากที่สุดเมื่อข้อมูลมีค่าสังเกต (n) ขนาดใหญ่ และตัวแปรอิสระมีความสัมพันธ์กันในระดับปานกลาง

2.3 การเลือกพารามิเตอร์การปรับแต่ง (Tuning Parameter) ด้วยวิธีการตรวจสอบไขว้ (Cross Validation)

การหาค่าพารามิเตอร์ปรับแต่ง (Tuning Parameter) มีความสำคัญอย่างมากในการวิเคราะห์การถดถอยเชิงเส้นด้วยวิธี Penalized Regression เพราะพารามิเตอร์ปรับแต่งมีผลต่อผลบวกกำลังสองของค่าความคลาดเคลื่อนสุ่ม ดังนั้น การเลือกพารามิเตอร์ปรับแต่งที่เหมาะสมจะทำให้ตัวแบบที่ได้มีผลบวกกำลังสองของค่าความคลาดเคลื่อนสุ่มต่ำ นั่นคือตัวแบบที่ได้มีความแม่นยำในการพยากรณ์สูงนั่นเอง

วิธีการตรวจสอบไขว้ (Cross Validation) เป็นวิธีการที่ง่ายและใช้กันอย่างแพร่หลายที่สุดในการประมาณค่าความคลาดเคลื่อนในการประมาณ (Prediction Error) (James et al., 2013) โดยวิธี Cross Validation มีหลายเทคนิค ประกอบด้วย 1) Leave-One-Out Cross Validation 2) Generalized Cross Validation 3) Leave-K-Out Cross Validation และ 4) K-Fold Cross Validation แต่ K-Fold Cross Validation เป็นเทคนิคที่ได้รับความนิยมเพราะเป็นวิธีที่สามารถลดความเอนเอียงของค่าประมาณและขณะเดียวกันก็มีค่าใช้จ่ายในการคำนวณไม่สูงนักเมื่อเปรียบเทียบกับ Cross Validation เทคนิคอื่น (Syed, 2011)

เทคนิค K-Fold Cross Validation เริ่มแรกจะทำการแบ่งข้อมูลออกเป็น 2 ส่วน คือ 1) ข้อมูลชุดฝึกฝน (Training Set) ใช้สร้างตัวแบบ และข้อมูลชุดตรวจสอบ (Test Set) ใช้ในการ

ตรวจสอบตัวแบบที่ถูกสร้างจากข้อมูลชุกฝึกฝน เทคนิค K-Fold Cross Validation จะนำข้อมูลชุดฝึกฝนมาแบ่งข้อมูลออกเป็น K ส่วนเท่า ๆ กัน จากนั้นนำข้อมูลจำนวน 1 ชุดมาเป็นข้อมูลชุดทดสอบย่อย (Sub-test Set) และนำข้อมูลที่เหลืออีก K-1 ชุดมาเป็นข้อมูลชุดฝึกฝนย่อย (Sub-training Set) เพื่อประมาณค่าพารามิเตอร์ที่สนใจจากข้อมูลชุดฝึกฝนย่อย จากนั้นนำตัวแบบที่ได้ไปทดสอบกับข้อมูลชุดทดสอบย่อย แล้วคำนวณหาค่าความคลาดในการพยากรณ์ (Prediction Error) ทำเช่นนี้ไปเรื่อยจนครบ K ครั้ง ซึ่งข้อมูลทั้ง K ชุดจะถูกนำมาใช้เป็นข้อมูลชุดฝึกฝนย่อยจนครบ จากนั้นนำความคลาดในการพยากรณ์ทั้ง K ครั้งมาหาค่าเฉลี่ย การทำ K-Fold Cross Validation มีกระบวนการโดยละเอียดดังนี้

1) กำหนดให้ $k:\{1,2,3,...,n\}\longmapsto\{1,2,3,...,K\}$ เป็นฟังก์ชันดัชนี (Indexing Function) ที่ใช้ในแบ่งค่าสังเกตที่ i (Observation i) ลงใน Fold ต่าง ๆ แบบสุ่มเป็นจำนวน K กลุ่ม กลุ่มละเท่า ๆ กัน

2) กำหนด λ เป็นตัวประมาณแบบจุดที่มีค่าแตกต่างกัน จะได้เซตของ λ ที่เป็นไปได้ประกอบด้วยสมาชิกของเซตจำนวน q ตัว ดังนั้น $\lambda=\lambda_1,\lambda_2,\lambda_3,...,\lambda_q$

3) สร้างตัวแบบครั้งที่ k โดยกำหนดข้อมูลจำนวน K-1 ชุดใช้ในการสร้างตัวแบบ และข้อมูลจำนวน 1 ชุด ชุดที่ k เป็นข้อมูลชุดทดสอบย่อย

- ในแต่ละ $\lambda_l : l=1,2,3,...,q$ ทำการสร้างตัวแบบ $\hat{f}^{-k}(x,\lambda)$ และประมาณค่าพารามิเตอร์ $\{\hat{\beta}_i\}_l$ แล้วนำไปหาค่าพยากรณ์ $\{\hat{y}_i\}_l$ ของข้อมูลชุดที่ k
- หาค่าความคลาดเคลื่อนในการพยากรณ์ (Prediction of Error) ของ λ_l จากสมการที่ (30)

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

$$e_k(\lambda_l)=\sum_{i=1}^{n_k}L(y_i,\{\hat{y}_i\}_l)^2 \tag{30}$$

4) ในแต่ละ $\lambda_l : l=1,2,3,...,q$ จะถูกนำมาสร้างตัวแบบจำนวน k ครั้ง ซึ่งค่าความคลาดเคลื่อนในการพยากรณ์ของ λ_l ทั้ง k ครั้งจะถูกนำมาค่าค่าเฉลี่ยความคลาดในการพยากรณ์ (Average Prediction of Error) ของ λ_l จากสมการที่ (31)

$$CV(\lambda_l)=\frac{1}{n}\sum_{k=1}^{K}e_k(\lambda_l) \tag{31}$$

4) ตัว $\hat{\lambda}$ ที่มีค่า $CV(\lambda_l)$ น้อยที่สุดจะเป็นพารามิเตอร์ปรับแต่งที่เหมาะสมที่สุด ซึ่งคำตอบของ $\hat{\lambda}$ เขียนได้ดังสมการที่ (32)

$$\hat{\lambda} = \underset{\lambda}{\arg\min} CV(\lambda_l) = \frac{1}{n} \sum_{k=1}^{K} e_k(\lambda_l) \quad (32)$$

5) นำค่าพารามิเตอร์ปรับแต่งที่เหมาะสมมาใช้ในการสร้างตัวแบบด้วยข้อมูลชุดฝึกฝนทั้งหมดอีกครั้ง และจะถือว่าตัวแบบนี้เป็นตัวแบบที่ดีที่สุด

ในการทำ K-Fold Cross Validation หากกำหนดค่า K มากเกินไปก็จะเกิดปัญหา Overfitting (High Variance) แต่ในทางกลับกันหากกำหนดค่า K น้อยเกินไปก็อาจเกิดปัญหา Underfitting (High Bias) ขึ้นอยู่กับจำนวนค่าสังเกตที่ใช้ในการสร้างตัวแบบ โดยทั่วไปแล้วจะแนะนำให้ใช้ K=5 หรือ K=10 (Breiman, 1992; Hastie et al., 2009)

2.4 งานวิจัยในอดีตที่เกี่ยวข้อง

จากการศึกษาค้นคว้างานวิจัยในอดีตที่เกี่ยวข้อง สามารถสรุปได้ดัง ตารางที่ 2.1

ตารางที่ 2.1 งานวิจัยในอดีตที่เกี่ยวข้อง

วัตถุประสงค์ของการศึกษา	สรุปผลการศึกษา
Best Subset Selection Via a Modern Optimization Lens (Bertsimas et al., 2016)	
เปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าทั้งหมด 4 วิธี ได้แก่ - MIO Best Subset Selection - Lasso - Stepwise - Sparsenet	การศึกษาได้มีการจำลองข้อมูลเพื่อใช้ในการทดลอง โดยมีการปรับจำนวนตัวแปรที่มีค่าสัมประสิทธิ์ไม่เท่ากับศูนย์ ขนาดของสัมประสิทธิ์ ค่า SNR ขนาดความสัมพันธ์ของตัวแปรอิสระ และจำนวนตัวแปรอิสระด้วย ผลการศึกษาพบว่า MIO Beast Subset Selection มีประสิทธิภาพโดดเด่นที่สุด ตัวแบบที่ได้มีความบางเบา มีตัวแปรน้อย อธิบายได้ง่าย อย่างไรก็ตามหากจะพิจารณาในด้านประสิทธิภาพของการพยากรณ์เพียงอย่างเดียว โดยรวมจะพบว่าวิธี L1 จะมีประสิทธิภาพในการพยากรณ์มากกว่าวิธี MIO Best Subset Selection แต่ตัวแบบจะมีจำนวนตัวแปรมากกว่าตัวแบบจากวิธี MIO Best Subset Selection หลายเท่าตัว
Extended Comparisons of Best Subset Selection, Forward Stepwise Selection, and the Lasso Following “Best Subset Selection from a Modern Optimization Lens” by Bertsimas, King, and Mazumder (2016) (Hastie et al., 2017)	
เปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าทั้งหมด 4 วิธี ได้แก่	การศึกษาได้จำลองข้อมูล โดยกำหนดให้ตัวแปรอิสระมีการแจกแจงปกติหลายตัวแปร ตัวแปรอิสระมีความสัมพันธ์กันปานกลาง (Rho =

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

วัตถุประสงค์ของการศึกษา	สรุปผลการศึกษา
- MIO Best Subset Selection (on Gurobi Solver) - Forward Stepwise Selection - L1 (Lasso) - Relaxed L1 (Relaxed Lasso)	0.35) กำหนดให้ SNR มีค่าตั้งแต่ 0.05 ถึง 6.00 และได้มีการปรับจำนวนตัวอย่าง จำนวนตัวแปรอิสระ และจำนวนตัวแปรอิสระที่มีค่าสัมประสิทธิ์การถดถอยไม่เท่ากับศูนย์เป็น 4 สถานการณ์ คือ - n = 100 p = 10 และ Nonzero entries = 5 - n = 500 p = 100 และ Nonzero entries = 5 - n = 50 p = 1000 และ Nonzero entries = 5 - n = 100 p = 1000 และ Nonzero entries = 10 ผลการทดลองพบว่า วิธี Relaxed Lasso มีประสิทธิภาพสูงที่สุดในทุกเมทริกซ์ของการวัดผล (Relative Risk, Relative Test Error, and Proportion of Variance Explained) และมีความเร็วในการคำนวณมากที่สุด ในกรณีที่ข้อมูลมีสัดส่วนของตัวแปรอิสระต่อจำนวนตัวอย่างสูง และมี SNR ต่ำ วิธี Lasso มีประสิทธิภาพสูงที่สุด ตามด้วยวิธี Best Subset Selection และ Forward Stepwise ตามลำดับ นอกจากนี้การศึกษายังพบว่าวิธี Best Subset Selection และ Forward Stepwise มีประสิทธิภาพใกล้เคียงกันแต่วิธี Forward Stepwise สามารถคำนวณได้รวดเร็วกว่า
Subset Selection with Shrinkage: Sparse Linear Modeling when the SNR is low (Mazumder et al., 2017)	
เปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าทั้งหมด 7 วิธี ได้แก่ - MIO-L0 (Best Subset Selection) - MIO-L0L1 (L1- regularized Best Subset Selection) - MIO-L0L2 (L2- regularized Best Subset Selection) - L1 (Lasso) - L1P (Polished version of the Lasso estimator) - L2 (Ridge Regression) - L1L2 (Elastic Net) - AL1L2 (Adaptive Elastic Net)	การศึกษาจำลองข้อมูล โดยกำหนดสถานการณ์ต่าง ๆ โดยมีการเปลี่ยนแปลงค่า SNR จำนวนตัวอย่าง จำนวนตัวแปรอิสระ และระดับความสัมพันธ์ของตัวแปรอิสระ (Rho) ให้มีค่าแตกต่างกัน ผลการทดลองพบว่า เมื่อระดับสัญญาณรบกวนสูง หรือ SNR มีค่าต่ำ (SNR = 1) วิธี L0 จะมีความแม่นยำในการพยากรณ์ต่ำ (Poorly Prediction Accuracy) และมักจะเกิดปัญหา Overfitting นอกจากนี้ยังเลือกตัวแปรเข้าสู่ตัวแบบน้อยเกินไป น้อยกว่าจำนวนตัวแปรที่มีค่าสัมประสิทธิ์ไม่เท่ากับศูนย์ที่แท้จริง (True Nonzero Coefficient) ส่วนวิธี L1 และวิธี L2 มีประสิทธิภาพในการพยากรณ์สูงกว่าวิธี MIO-L0 แต่มักจะเลือกตัวแปรเข้าสู่ตัวแบบมาก ส่วนวิธี L1P แม้ว่าจะเลือกตัวแปรเข้าสู่ตัวแบบน้อยกว่าวิธี L1 แต่กลับมีประสิทธิภาพในการพยากรณ์ต่ำกว่า ส่วนวิธี MIO-L0L1 และ MIO-L0L2 พบว่าทั้งสองวิธีมีประสิทธิภาพสูงในการพยากรณ์เมื่อเปรียบเทียบกับวิธีอื่น ๆ และยังสามารถแก้ปัญหา Overfitting ที่เกิดขึ้นกับวิธี L0 ได้เป็นอย่างดี แต่ตัวแบบที่ได้จากวิธี MIO-L0L1 และ MIO-L0L2 จะมีความ

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

วัตถุประสงค์ของการศึกษา	สรุปผลการศึกษา
	บางเบาน้อยกว่าตัวแบบจากวิธี MIO-L0 เมื่อ SNR มีค่าสูงขึ้น พบว่าวิธี MIO-L0, MIO-L0L1 และ MIO-L0L2 มีประสิทธิภาพใกล้เคียงกันทั้งด้านความบางเบาของตัวแบบและความแม่นยำในการพยากรณ์ และเมื่อเปรียบเทียบกับตัวแบบที่ดีที่สุดของวิธี L1 พบว่าแม้ว่าวิธี L1 จะมีความแม่นยำในการพยากรณ์สูงกว่า แต่ตัวแบบที่ได้กลับมีจำนวนตัวแปรมากกว่าตัวแบบจากวิธี MIO-L0, MIO-L0L1 และ MIO-L0L2 เป็นจำนวนมาก
Fast Best Subset Selection: Coordinate Descent and Local Combinatorial Optimization Algorithms (Hazimeh & Mazumder, 2020)	
เปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าทั้งหมด 8 วิธี ได้แก่ - L0Learn (Best Subset Selection) - L0L1Learn (L1- regularized Best Subset Selection) - L0L2Learn (L2- regularized Best Subset Selection) - Forward Stepwise Selection - L1 (Lasso) - L1Relaxed (Relaxed Lasso) - L1L2 (Elastic Net) - L2 (Ridge Regression) - MCP (Minimax Concave Penalty) - IHT (Iterative Hard Thresholding)	การศึกษาได้มีการจำลองข้อมูล โดยกำหนดให้ตัวแปรอิสระมีการแจกแจงปกติหลายตัวแปร และเวกเตอร์ความคลาดเคลื่อนที่ใช้ในการจำลองเวกเตอร์ของตัวแปรตามจะแปรผันไปตามระดับ SNR และกำหนดให้ค่าสัมประสิทธิ์ของตัวแปรที่มีค่าไม่เท่ากับศูนย์มีค่าเท่ากับ 1 ผลการทดลองพบว่า ในกรณีที่ตัวแปรอิสระมีความสัมพันธ์สูง (Rho = 0.9) วิธี L0L2Learn มีประสิทธิภาพสูงที่สุดทั้งในด้านของความน่าจะเป็นในการเรียกคืนตัวแบบ (Recovery Probability) ขนาดของตัวแบบที่มีขนาดเล็กกระทัดรัด (Support Size) และความคลาดเคลื่อนในการพยากรณ์ (Prediction Error) แม้ว่าขนาดของตัวอย่างจะมีการเปลี่ยนแปลง แต่เมื่อมีขนาดตัวอย่างมากกว่า 300 ตัวอย่าง วิธี L0Learn จะมีประสิทธิภาพเทียบเคียงกับวิธี L0L2Learn ในกรณีที่ตัวแปรอิสระมีความสัมพันธ์ปานกลาง (Rho = 0.5) พบว่าวิธี L0Learn และ L0L2Learn มีประสิทธิภาพใกล้เคียงกันและมีประสิทธิภาพสูงเมื่อเทียบกับวิธีอื่น ๆ เมื่อจำนวนตัวอย่าง 300 ตัวอย่าง โดยประมาณ ทั้งสองวิธีจะสามารถกู้คืนตัวแบบได้อย่างสมบูรณ์ (Fully Recovery) แต่หากเปรียบเทียบผลลัพธ์เมื่อมีการเปลี่ยนแปลงระดับ SNR พบว่าในภาพรวมพบว่าวิธี L0L2Learn และวิธี L1L2 มีประสิทธิภาพสูงใกล้เคียงกัน แต่วิธี L1L2 จะให้ตัวแบบที่มีจำนวนตัวแปรมากกว่า และอาจจะมีตัวแปรในตัวแบบมากถึง 90% ของตัวแปรอิสระเมื่อ SNR = 1 ในกรณีที่ SNR สูง (SNR = 100) วิธี L0L2Learn มีประสิทธิภาพสูงที่สุด สามารถกู้คืนตัวแบบได้อย่างสมบูรณ์ (Fully Recovery) แต่เมื่อ SNR ต่ำวิธี Ridge Regression มีประสิทธิภาพด้านการพยากรณ์สูง (มีค่า Prediction error ต่ำ) กว่าวิธี L0L2Learn แต่วิธี L0L2Learn จะให้ตัว

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

วัตถุประสงค์ของการศึกษา	สรุปผลการศึกษา
	แบบที่บางเบากว่า ส่วนวิธี L0Learn มีแนวโน้มจะเกิด Overfitting อย่างรวดเร็วเมื่อ SNR มีค่าต่ำ
Discussion of “Best Subset, Forward Stepwise or Lasso? Analysis and Recommendations Based on Extensive Comparisons” (Mazumder, 2020)	
เปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าทั้งหมด 6 วิธี ได้แก่ - L0Learn (Best Subset Selection) - L0L2Learn (L2- regularized Best Subset Selection) - Forward Stepwise Selection - L1 (Lasso) - L1Relaxed (Relaxed Lasso) - L2 (Ridge Regression)	การศึกษามีการจำลองข้อมูล และมีการรับ SNR ให้มีระดับต่าง ๆ กันตั้งแต่ 0.001 ไปจนถึง 100 และกำหนดให้ตัวแปรอิสระมีความสัมพันธ์แบบยกกำลัง (Exponential Correlation) เท่ากับ 0.5 ผลการทดลองพบว่า วิธี L0Learn จะให้ตัวแบบที่มีประสิทธิภาพในการพยากรณ์ต่ำแม้ว่า SNR จะอยู่ในระดับปานกลาง เช่นเดียวกับวิธี Forward Stepwise แต่การทำ Regularized-Subsets (L0L2Learn) สามารถเพิ่มประสิทธิภาพในการพยากรณ์ได้ผ่านการใช้ L2 shrinkage นอกจากนี้ยังพบว่าวิธี L0L2Learn จะให้ตัวแบบที่มีประสิทธิภาพการพยากรณ์ดีกว่าวิธี Relaxed Lasso อย่างไรก็ตาม เมื่อ SNR มีระดับต่ำมาก ๆ หรือข้อมูลมีสัญญาณรบกวนสูง วิธี L2 หรือ Ridge Regression จะมีประสิทธิภาพในการพยากรณ์มากที่สุด แต่วิธี L0L2Learn ก็มีประสิทธิภาพสูงเกือบเทียบเท่าวิธี L2 อีกทั้งตัวแบบที่ได้มีความบางเบามากกว่า ส่วนวิธี Relaxed Lasso นั้นแม้จะได้ตัวแบบที่มีตัวแปรน้อยกว่า (บางเบากว่า) แต่ได้ประสิทธิภาพในการพยากรณ์ต่ำกว่าวิธี L0L2Learn นอกจากนี้การศึกษานี้ยังได้ทดลองใช้กับชุดข้อมูลจริงทั้งหมด 3 ชุดข้อมูล (3 Data Set) ในภาพรวมพบว่าวิธี L1 จะได้ตัวแบบที่มี MSE ต่ำกว่าเพียงเล็กน้อย แต่กลับพบว่าตัวแบบมีขนาดใหญ่กว่าวิธี L0L2Learn และวิธี L0L1Learn หลายสิบเท่าตัว
Selecting Massive Variables Using an Iterated Conditional Modes/Medians Algorithm (Vitara Pungpapong et al., 2015)	
เปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าทั้งหมด 6 วิธี ได้แก่ - EBVS (Empirical Bayes Variable Selection) - L1 (Lasso) - AL1 (Adaptive Lasso) - SL1 (Scaled Lasso)	การศึกษาใช้ข้อมูลที่ได้จากการจำลอง โดยกำหนดให้ p= 1000, n = 100, Nonzero entries = 20 และตัวแปรอสระมีความสัมพันธ์แบบยกกำลัง (Exponential correlation) = 0, 0.1, 0.2, ..., 0.9 ผลการศึกษาพบว่า วิธี EBVS มีประสิทธิภาพดีกว่าวิธีอื่น ๆ ในด้านความสามารถในการพยากรณ์ (Prediction Performance) โดยเฉพาะอย่างยิ่งเมื่อ Rho มีค่าตั้งแต่ 0.3 วิธี EBVS จะมีความคงเส้นคงวาในการพยากรณ์ ส่วนวิธี L1 และวิธี Adaptive Lasso พบว่า วิธี L1 จะมี Prediction

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

วัตถุประสงค์ของการศึกษา	สรุปผลการศึกษา
	Performance สูงกว่าเมื่อ Rho > 0.3 แต่หาก Rho < 0.3 วิธี Adaptive Lasso จะมี Prediction Error ต่ำกว่า
การเปรียบเทียบการประมาณค่าพารามิเตอร์ของการวิเคราะห์การถดถอยที่ปรับด้วยฟังก์ชันลงโทษภายใต้ข้อมูลที่มีมิติสูง (เบญจมาส รุ่งศรานนท์ & อัชฌา อระวีพร, 2019)	
เปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าสัมประสิทธิ์การถดถอยที่ปรับด้วย Penalty function 5 วิธี ได้แก่ - L2 (Ridge) - L1 (Lasso) - L1L2 (Elastic Net) - A-L1 (Adaptive Lasso) - A-L1L2 (Adaptive Elastic Net)	ในการศึกษาได้จำลองข้อมูลสูงมิติโดยกำหนดให้ตัวแปรอิสระมีการแจกแจงแบบปกติ แต่มีการจำลองค่าความคลาดเคลื่อนของตัวแบบทั้งหมด 3 รูปแบบ คือ 1) การแจกแจงปกติ 2) การแจกแจงปกติปลอมปน และการแจกแจงไวบูล และจากผลการทดลองพบว่า A-L1L2 ให้ค่าเฉลี่ยความคลาดเคลื่อนกำลังกำลังสองเฉลี่ยที่ต่ำที่สุดในทุกกรณี นอกจากนี้งานวิจัยยังนำทั้ง 5 วิธีมาประยุกต์ใช้กับข้อมูลจริงที่มีขนาดตัวอย่าง 5 10 และ 15 ตัวอย่าง จำนวนตัวแปรอิสระ 16 ตัว ผลปรากฏว่าวิธี A-L1L2 เป็นวิธีการที่ดีกว่าวิธีอื่น ๆ เช่นเดียวกับผลที่ได้จากข้อมูลจำลอง นอกจากนี้ยังพบว่าเมื่อข้อมูลมีขนานตัวอย่างมากขึ้น ค่าเฉลี่ยของความคลาดเคลื่อนกำลังสองที่ได้จากวิธี A-L1L2 จะมีค่าต่ำลง
การเปรียบเทียบประสิทธิภาพการประมาณค่าพารามิเตอร์ด้วยวิธีแลซโซในการถอดถอยเชิงเส้นที่มีมิติสูง (พัชราภรณ์ พรดำเนินสวัสดิ์, 2560)	
เปรียบเทียบประสิทธิภาพของตัวประมาณค่าสัมประสิทธิ์การถดถอยเชิงเส้น ทั้งหมด 4 วิธี ได้แก่ - L2 (Ridge) - L1 (Lasso) - Relaxed Lasso - A-L1 (Adaptive Lasso)	ในการศึกษาได้จำลองข้อมูล โดยมุ่งเน้นศึกษาประสิทธิภาพในกรณีที่ข้อมูลมีเมทริกซ์สหสัมพันธ์ของตัวแปรอิสระมีความแตกต่างกัน ผลการศึกษาพบว่า กรณีที่เป็นตัวแบบถดถอยบางเบา เมทริกซ์สหสัมพันธ์ของตัวแปรอิสระเป็นเชิงตั้งฉากกัน และโครงสร้างแบบกลุ่ม วิธี Relaxed Lasso มีประสิทธิภาพด้านการพยากรณ์และประสิทธิภาพด้านการประมาณสูงที่สุด วิธี A-L1 มีประสิทธิภาพสูงกว่าวิธี L1 เมื่อตัวแปรอิสระมีความสัมพันธ์กันสูง และมีขนาดตัวอย่างจำนวนมาก กรณีที่ไม่เป็นตัวแบบถดถอยบางเบา เมทริกซ์สหสัมพันธ์ของตัวแปรอิสระเป็นไม่เชิงตั้งฉากกัน วิธี L2 มีประสิทธิภาพด้านการพยากรณ์และประสิทธิภาพด้านการประมาณสูงที่สุด รองลงมาคือวิธี L1 ส่วนวิธี A-L1 จะมีประสิทธิภาพดีกว่าวิธี Relaxed Lasso เมื่อตัวแปรอิสระมีความสัมพันธ์กันสูง
การเปรียบเทียบวิธีคัดกรองตัวแปรสำหรับการทดสอบกลุ่มของสัมประสิทธิ์การถดถอยที่มีมิติสูงแบบเป็นนำดับขั้น (สวรรยา ภู่เงิน, 2557)	
เปรียบเทียบวิธีการคัดกรองตัวแปรอิสระด้วยวิธี - L1 (Lasso)	การศึกษาใช้การจำลองข้อมูลโดยกำหนดให้จำนวนตัวอย่างต่อจำนวนตัวแปรเท่ากับ 100:500 และ 100:1000 และกำหนดความสัมพันธ์ของตัวแปรอิสระเป็น 0, 0.5, และ 0.9 นอกจากนี้ยังมีการ

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

วัตถุประสงค์ของการศึกษา	สรุปผลการศึกษา
- A-L1 (Adaptive Lasso) - L1L2 (Elastic Net)	ทดลองในชุดข้อมูลจริง 2 ชุด ซึ่งเป็นชุดข้อมูลที่ตัวแปรอิสระมีความสัมพันธ์แบบปกติและมีความสัมพันธ์กันสูง ผลการศึกษาพบว่า วิธี L1 มีความถูกต้องในการคัดเลือกตัวแปรเข้าสู่ตัวแบบมากที่สุดในกรณีที่ตัวแปรอิสระมีความสัมพันธ์เท่ากับ 0.5 และ 0.9 แต่หากตัวแปรอิสระมีความสัมพันธ์เท่ากับ 0.0 ทั้ง 3 วิธีมีประสิทธิภาพเท่ากัน แต่เมื่อพิจารณาถึงอัตราส่วนของการเกิด Type 1 Error กลับพบว่าวิธี A-L1 และ L1L2 มีประสิทธิภาพดีกว่าวิธี L1

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

บทที่ 3
วิธีดำเนินการวิจัย

การวิจัยครั้งนี้เป็นการวิจัยเชิงทดลองโดยวิธีการจำลอง (Simulation Study) โดยมีวัตถุประสงค์เพื่อศึกษาพฤติกรรม และเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าสัมประสิทธิ์การถดถอยเชิงเส้นสำหรับข้อมูลที่มีมิติสูง ด้วยวิธี Penalized Regression วิธี L0Learn และวิธี L0L2Learn เปรียบเทียบกับวิธี Penalized Regression ที่ได้รับความนิยมในปัจจุบัน ได้แก่ วิธี L1 วิธี A-L1 และ วิธี A-L1L2 โดยพิจารณาจากความถูกต้องของการประมาณค่าสัมประสิทธิ์การถดถอยหรือประสิทธิภาพด้านการพยากรณ์จะวัดจากค่าเฉลี่ยความคลาดเคลื่อนการทำนาย (Average of Mean Squared Error: AMSE) และความถูกต้องในการคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบ โดยพิจารณาจากค่าเฉลี่ยของค่า Precision ค่าเฉลี่ยของค่า Recall และค่าเฉลี่ยของค่า AUC โดยการวิจัยครั้งนี้ใช้โปรแกรม R ในการสร้างข้อมูลจำลอง (Simulation Data) วิเคราะห์ตัวแบบ รวมถึงเปรียบเทียบประสิทธิภาพของการวิเคราะห์ตามสถานการณ์ที่กำหนดตามขอบเขตของการวิจัย ซึ่งในงานวิจัยมีขอบเขตและขั้นตอนดังนี้

3.1 ขอบเขตของการวิจัยและการจำลองข้อมูล

ในการวิจัยได้กำหนดสถานการณ์ในการจำลองข้อมูล ดังนี้

- จำนวนตัวอย่างเท่ากับ 100 ตัวอย่าง ($n = 100$)
- จำนวนตัวแปรอิสระในตัวแบบเท่ากับ 1000 ตัว ($p = 1000$)
- ตัวแปรอิสระทุกตัวมีความสัมพันธ์กัน 3 ระดับ คือ 0.0, 0.5 และ 0.9 ($\rho = 0.0, 0.5, 0.9$)
- กำหนดอัตราส่วนสัญญาณต่อสัญญาณรบกวนทั้งหมด 5 ระดับ คือ 0.1, 0.5, 1, 5, 10 และ 20 ($\text{SNR} = 0.1, 0.5, 1, 5, 10, 20$)
- X เป็นเมทริกซ์ตัวแปรอิสระขนาด $n \times p$ และ $X_i = N(0, \Sigma)$ โดยที่ Σ เป็นเมทริกซ์ความแปรปรวนร่วมที่ $\Sigma_{i,j} = \rho^{|i-j|}$ เมื่อ $i, j = 1,2,3,...,p$ จะได้ว่า

$$\Sigma_{i,j} = \rho^{|i-j|} = \begin{bmatrix} \rho^{|1-1|} & \cdots & \rho^{|1-p|} \\ \vdots & \ddots & \vdots \\ \rho^{|p-1|} & \cdots & \rho^{|p-P|} \end{bmatrix} \quad (33)$$

- β เป็นเวกเตอร์ค่าสัมประสิทธิ์การถดถอยขนาด $p \times 1$ ซึ่งมีสัมประสิทธิ์การถดถอยไม่เท่ากับศูนย์ 20 ตัว คือ $\beta_1, ..., \beta_{10} = 2$ และ $\beta_{101}, ..., \beta_{110} = 1$

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

- ε คือ เวกเตอร์ความคลาดเคลื่อนขนาด $n \times 1$ และ $\varepsilon \sim N(0, \sigma^2)$ โดยที่ σ^2 ขึ้นอยู่กับค่า SNR ซึ่งคำนวณได้ดังสมการ (34)

$$SNR = \frac{\beta^T \Sigma \beta}{\sigma^2} \tag{34}$$

- Y เป็นเวกเตอร์ตัวแปรตามขนาด $n \times 1$ ซึ่งสร้างจากตัวแบบดังสมการที่ (1)
- จำลองข้อมูล 100 รอบในแต่ละสถานการณ์

โดยหลักเกณฑ์ในการกำหนดค่าพารามิเตอร์ต่าง ๆ ที่ใช้ในการจำลองข้อมูล ยึดตามการศึกษาของ Hazimeh และ Mazumder (Hazimeh & Mazumder, 2020) และ Hastie และคณะ (Hastie et al., 2017) ซึ่งเปรียบเทียบประสิทธิภาพการคัดเลือกตัวแปรด้วยวิธี L0Learn, L0L2Learn, L1 และ A-L1 และการศึกษาของ Pungpapong และคณะ (V. Pungpapong et al., 2015) ซึ่งศึกษาการคัดเลือกตัวแปรด้วยวิธีเบส์ (Bayesian Approach) ซึ่งจะทำให้เปรียบเทียบประสิทธิภาพของการคัดเลือกตัวแปรได้หลายวิธีมากขึ้น สามารถอภิปรายความเหมือน ความแตกต่างของผลการศึกษาได้ นอกจากนี้ในการกำหนดสัมประสิทธิ์การถดถอยไม่เท่ากับศูนย์เป็น Block ถือเป็นปัญหาที่ยากปัญหาหนึ่งเรียกว่า Hard Sparsity

3.2 เกณฑ์ที่ใช้ในการพิจารณา

เกณฑ์ที่ใช้ในการพิจารณาเปรียบเทียบประสิทธิภาพของการประมาณค่าสัมประสิทธิ์การถดถอยด้วยวิธี L0Learn, วิธี L0L2Learn, วิธี L1, วิธี A-L1, และวิธี A-L1L2 ได้แก่

Chulalongkorn University

3.2.1. ประสิทธิภาพของการพยากรณ์

การวัดประสิทธิภาพของการพยากรณ์ ใช้บ่งบอกว่าค่าพยากรณ์ที่ได้จากการวิเคราะห์ใกล้เคียงกับค่าสังเกตมากเพียงใด วัดจากค่าเฉลี่ยความคลาดเคลื่อนกำลังสอง (Average of Mean Squared Error: AMSE) ที่ได้จากการจำลองข้อมูลทั้งหมด 100 ครั้ง ซึ่งมีสูตรในการคำนวณตามสมการ (35)

$$AMSE = \frac{\sum_{r=1}^{100} MSE_r}{100} \tag{35}$$

เมื่อ $r = 1,2,3,...,100$ หรือเท่ากับรอบในการจำลองข้อมูล

โดยที่ค่า Mean Squared Error สามารถคำนวณได้ตามสมการที่ (36)

$$MSE_r = \frac{\sum_{r=1}^{n} (y_i - \hat{y}_i)^2}{n} \quad (36)$$

3.2.2. ประสิทธิภาพในการเลือกตัวประมาณ

นอกจากประสิทธิภาพของการพยากรณ์แล้ว อีกประเด็นหนึ่งที่ต้องให้ความสำคัญในการวิเคราะห์ตัวแบบบางเบา (Sparse Model) คือความสามารถของตัวแบบในการคัดเลือกตัวแปรอิสระที่ถูกต้องเข้าสู่ตัวแบบ

ความผิดพลาดในการคัดเลือกตัวแปรมี 2 แบบคือ 1) กรณีที่ค่าสัมประสิทธิ์ไม่เท่ากับศูนย์ (True Non-zero Coefficient) แต่ตัวแปรอิสระไม่ถูกคัดเลือกเข้าสู่ตัวแบบหรือมีค่าสัมประสิทธิ์จากการประมาณมีค่าเท่ากับศูนย์ และ 2) กรณีที่ค่าสัมประสิทธิ์เท่ากับศูนย์ (True Zero Coefficient) แต่ตัวแปรอิสระถูกคัดเลือกเข้าสู่ตัวแบบหรือมีค่าสัมประสิทธิ์จากการประมาณมีค่าไม่เท่ากับศูนย์ ซึ่งสามารถอธิบายให้เข้าใจได้ง่ายด้วยเมทริกซ์ความสับสน (Confusion Matrix)

		Predicted	
		Non-zero Coefficient	Zero Coefficient
Actual	Non-zero Coefficient	True Positive	False Negative
	Zero Coefficient	False Positive	True Negative

สำหรับประสิทธิภาพในการคัดเลือกตัวประมาณเข้าสู่ตัวแบบจะพิจารณาจาก 3 ตัวชี้วัด คือ

1) ค่าเฉลี่ยของค่าความแม่นยำ (Mean of Precision) เป็นการวัดความแม่นยำในการคัดเลือกตัวแปร คือ โดยเฉลี่ยแล้วตัวแบบสามารถคัดเลือกตัวแปรที่ถูกต้องได้เป็นสัดส่วนเท่าใดเมื่อเทียบกับตัวแปรที่ถูกคัดเลือกเข้าสู่ตัวแบบทั้งหมด กล่าวได้อีกอย่างหนึ่งคือตัวแบบสามารถคัดเลือกตัวแปรที่มีค่าสัมประสิทธิ์ไม่เท่ากับศูนย์ในชุดข้อมูลจริง (True Non-zero Coefficient) เป็นสัดส่วนเท่าใดเมื่อเปรียบเทียบกับจำนวนตัวแปรที่มีค่าสัมประสิทธิ์ไม่เท่ากับศูนย์ทั้งหมดในตัวแบบ (Non-zero Coefficient) ซึ่งมีสูตรการคำนวณดังสมการ (37)

$$Average\ of\ Precision = \frac{\sum_{r=1}^{100} Precision_r}{100} \quad (37)$$

เมื่อ $r = 1,2,3,...,100$ หรือเท่ากับรอบในการจำลองข้อมูล

โดยที่ค่า Precision สามารถคำนวณได้ตามสมการที่ (38)

$$Precision_r = \frac{True\ Positive}{True\ Positive + False\ Positive} \quad (38)$$

เมื่อ True Positive คือ จำนวนตัวแปรที่มีค่าสัมประสิทธิ์ไม่เท่ากับศูนย์ในตัวแบบ และมีค่าสัมประสิทธิ์ไม่เท่ากับศูนย์ในชุดข้อมูลจริง

False Positive คือ จำนวนตัวแปรที่มีค่าสัมประสิทธิ์เท่ากับศูนย์ในชุดข้อมูลจริง แต่มีค่าสัมประสิทธิ์ไม่เท่ากับศูนย์ในตัวแบบ

2) ค่าเฉลี่ยของค่าความระลึก (Mean of Recall หรือ Mean of Sensitivity) เป็นการวัดความสามารถในการเรียกคืนตัวแปร คือ โดยเฉลี่ยแล้วตัวแบบสามารถคัดเลือกตัวแปรที่ถูกต้องได้เป็นสัดส่วนเท่าใดเมื่อเทียบกับตัวแปรที่ถูกต้องทั้งหมด กล่าวได้อีกอย่างหนึ่งคือตัวแบบสามารถเรียกคืนตัวแปรที่เป็นมีค่าสัมประสิทธิ์ไม่เท่ากับศูนย์ในชุดข้อมูลจริง (True Non-zero Coefficient) คิดเป็นสัดส่วนเท่าใดเมื่อเทียบกับจำนวนตัวแปรที่เป็นมีค่าสัมประสิทธิ์ไม่เท่ากับศูนย์ในชุดข้อมูลจริง (True Non-zero Coefficient) ซึ่งในที่นี้มี True Non-zero Coefficient ทั้งหมด 20 ตัว ซึ่งมีสูตรการคำนวณดังสมการ (39)

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

$$Average\ of\ Recall = \frac{\sum_{r=1}^{100} Recall_r}{100} \quad (39)$$

เมื่อ $r = 1,2,3,...,100$ หรือเท่ากับรอบในการจำลองข้อมูล

โดยที่ค่า Recall สามารถคำนวณได้ตามสมการที่ (40)

$$Recall_r = \frac{True\ Positive}{True\ Positive + False\ Negative} \quad (40)$$

เมื่อ True Positive คือ จำนวนตัวแปรที่มีค่าสัมประสิทธิ์ไม่เท่ากับศูนย์ในตัวแบบ และมีค่าสัมประสิทธิ์ไม่เท่ากับศูนย์ในชุดข้อมูลจริง

False Negative คือ จำนวนตัวแปรที่มีค่าสัมประสิทธิ์ไม่เท่ากับศูนย์ในชุดข้อมูลจริง แต่มีค่าสัมประสิทธิ์เท่ากับศูนย์ในตัวแบบ

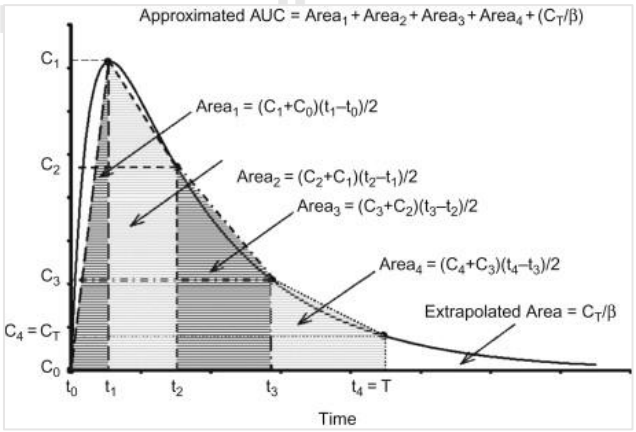
3) ค่าเฉลี่ยของค่าพื้นที่ใต้กราฟ ROC (Mean of Receiver Operating Characteristic) หรือค่าเฉลี่ยของค่า AUC (Area Under the Curve) ใช้ระบุความน่าจะเป็นว่าตัวแบบสามารถแยกตัวแปรที่เป็น True Non-zero Coefficient และ True Zero Coefficient ออกจากกันได้ดีเพียงใด

ROC Curve เกิดจากการสร้างกราฟความสัมพันธ์ระหว่าง True Positive Rate (Sensitivity) และ False Positive Rate (1–Specificity) โดย ROC Curve ที่ต้องการคือ ROC Curve ที่มีค่า True Positive Rate และค่า False Positive Rate สูง ซึ่งจะบ่งบอกว่าตัวแบบสามารถแยกตัวแปรที่เป็น True Non-zero Coefficient และ True Zero Coefficient ออกจากกันได้ดีนั่นเอง

ค่า AUC หาได้จากพื้นที่ใต้กราฟ (Area Under the Curve: AUC) โดยใช้หลักของการคำนวณพื้นที่สี่เหลี่ยมคางหมู (Trapezoidal Rule) เข้ามาช่วย โดยการย่อยพื้นที่ใต้กราฟเป็นสี่เหลี่ยมคางหมูหลาย ๆ รูป คำนวณหาพื้นที่สี่เหลี่ยมคางหมูย่อยทุกรูป แล้วจึงนำมารวมกัน จะได้เป็นพื้นที่ใต้กราฟทั้งหมด โดยเขียนเป็นสมการได้ดังสมการที่ (41)

$$AUC = \sum_{i=1}^{n-1} \frac{1}{2}(C_i + C_{i+1})(t_{i+1} - t_i) \tag{41}$$

หรือสามารถอธิบายได้ดังรูปที่ 3.1



รูปที่ 3.1 การคำนวณพื้นที่ใต้กราฟโดยใช้หลักสี่เหลี่ยมคางหมู (Baynes et al., 2012)

3.3 วิธีการดำเนินงานวิจัย

ในการดำเนินการวิเคราะห์สามารถทำได้ตามขั้นตอนดังนี้

3.3.1 ขั้นตอนการสร้างข้อมูล

สร้าง (Generate) ข้อมูลค่าสังเกต $(Y_i, X_i);\ i = 1,2,3,...,n$ ของชุดข้อมูลอย่างละ n ค่า หรือ 100 ตัว โดยมีขั้นตอนในการสร้างโดยละเอียด ดังนี้

1) สร้างเมทริกซ์ของตัวแปรอิสระ X ขนาด $n \times p$ โดยที่ X_i มีการแจกแจงปกติหลายตัวแปร (Multivariate Normal Distribution) ที่มีเวกเตอร์ค่าเฉลี่ยเท่ากับเวกเตอร์ศูนย์ และเมทริกซ์ความแปรปรวนร่วมเท่ากับ Σ นั่นคือ $X_i \sim N(0,\Sigma);\ i = 1,2,3,...,n$ ซึ่งเมทริกซ์ความแปรปรวนร่วมจะถูกกำหนดตามสถานการณ์ต่าง ๆ ตามขอบเขตของการวิจัยที่ระบุในหัวข้อ 3.1

2) หาค่าความแปรปรวนสำหรับใช้สร้างเวกเตอร์ความคลาดเคลื่อนสุ่มซึ่งคำนวณตามสมการที่ (42)

$$\sigma^2 = \frac{\beta^T \Sigma \beta}{SNR} \tag{42}$$

โดยค่าความแปรปรวนจะถูกกำหนดตามค่า SNR ตามสถานการณ์ต่าง ๆ ที่ได้ระบุตามขอบเขตของการวิจัย

3) สร้างเวกเตอร์ความคลาดเคลื่อนสุ่ม ε ขนาด $n \times 1$ โดยที่ ε_i มีการแจกแจงปกติที่มีค่าเฉลี่ยเท่ากับศูนย์และค่าความแปรปรวนคงที่สำหรับ $i = 1,2,3,...,n$ และกำหนดส่วนเบี่ยงเบนมาตรฐานเท่ากับรากที่สองของค่าความแปรปรวนที่ได้จากขั้นตอนที่ 1.2

4) สร้างเวกเตอร์ของตัวแปรตาม Y จากตัวแบบในสมการที่ (1) เมื่อ X และ ε เป็นข้อมูลที่ได้จากขั้นตอนที่ 1) และ 3) ตามลำดับ และกำหนดเวกเตอร์ของค่าสัมประสิทธิ์การถดถอย β ที่ได้ระบุตามขอบเขตของการวิจัย

3.3.2 ขั้นตอนการหาตัวประมาณด้วยวิธี L0Learn

1) นำชุดข้อมูลตัวแปรอิสระ X และตัวแปรตาม Y ที่ได้จากขั้นตอนที่ 3.3.1 มาหาพารามิเตอร์ปรับแต่ง (Tunning Parameter) ด้วยวิธี K-fold Cross Validation โดยในการศึกษาจะกำหนดให้ $K = 5$ และใช้ค่าความคลาดเคลื่อนกำลังสองเฉลี่ย (Mean Squared Error: MSE) เป็นตัววัดประสิทธิภาพของพารามิเตอร์ปรับแต่ง

1.1) แบ่งข้อมูลออกเป็น 5 ส่วนเท่า ๆ กัน จะได้ T_1, T_2, T_3, T_4, T_5 ให้ข้อมูล 1 ส่วนเป็นชุดข้อมูลทดสอบย่อย และข้อมูลอีก 4 ชุดเป็นข้อมูลฝึกฝนย่อย

1.2) กำหนดค่าพารามิเตอร์ปรับแต่ง λ_2^j ให้มีค่าเข้าใกล้ศูนย์มากที่สุด โดยปกติค่าพื้นฐานจะกำหนดให้ λ_2^j = 0.0001 และกำหนดค่าพารามิเตอร์ปรับแต่ง $\lambda_0^j > 0$ เมื่อ λ_0^j เป็นค่าคงที่ใด ๆ ขึ้นอยู่กับลักษณะของข้อมูล ซึ่งจะมีจำนวนทั้งหมด 100 ค่า จากนั้น ณ จุด λ_0^j ใด ๆ ทำการวิเคราะห์หาตัวแบบที่ทำให้ MSE มีค่าต่ำที่สุด

1.3) จากนั้นนำตัวแบบที่ได้จากขั้นตอนที่ 1.2 ไปใช้ในการพยากรณ์ค่าตัวแปรตาม Y ในชุดข้อมูลทดสอบย่อยแล้วคำนวณค่า MSE ของแต่ละชุด ทำเช่นนี้จนครบ 5 ชุดจะได้ค่า MSE ทั้งหมด 5 ค่า จากนั้นหาค่าเฉลี่ยของค่า MSE ที่จุด λ_0^j ใด ๆ

1.4) เลือก λ_0^j ที่ทำให้ค่าเฉลี่ยของค่า MSE มีค่าต่ำที่สุด จะได้ว่าค่า λ_0^j นั้นคือพารามิเตอร์ปรับแต่งที่เหมาะสมที่สุด

2) นำค่าพารามิเตอร์ปรับแต่งที่ได้จากข้อ 1 มาหาค่าสัมประสิทธิ์การถดถอยจากข้อมูลทั้งหมดอีกครั้ง

3) จากนั้นนำตัวแบบที่เหมาะสมที่สุดที่ฝึกฝนด้วยข้อมูลชุดที่ 1 ที่ได้จากขั้นตอนที่ 2 มาคำนวณค่า MSE จากชุดทดสอบ โดยกำหนดให้ข้อมูลชุดที่ 2 ที่ถูกจำลองด้วยเงื่อนไขเดียวกันเป็นข้อมูลชุดทดสอบเพื่อหาค่า MSE และเมื่อใช้ข้อมูลชุดที่ 2 ซึ่งเป็นข้อมูลชุดฝึกฝนในการหาค่า MSE จะกำหนดให้ข้อมูลชุดที่ 3 ที่ถูกจำลองด้วยเงื่อนไขเดียวกันเป็นข้อมูลชุดทดสอบเพื่อหาค่า MSE ทำเช่นนี้ไปเรื่อย ๆ จนตัวแบบที่ได้จากการฝึกฝนด้วยข้อมูลชุดที่ 100 จะใช้ข้อมูลชุดที่ 1 เป็นข้อมูลชุดทดสอบในการหาค่า MSE

4) นับจำนวนสัมประสิทธิ์การถดถอยที่เป็น True Positive, False Negative, False Positive, และ True Negative จากนั้นคำนวณหาค่า Precision, Recall, และ AUC ตามลำดับ

5) ทำซ้ำขั้นตอนที่ 1 ถึง 5 จำนวน 100 ครั้งตามการจำลองข้อมูล 100 รอบในแต่ละสถานการณ์

6) หาค่าเฉลี่ยของค่า MSE, Precision, Recall, และ AUC จากผลการทดลองทั้ง 100 ครั้ง

3.3.3 ขั้นตอนการหาตัวประมาณด้วยวิธี L0L2Learn

1) นำชุดข้อมูลตัวแปรอิสระ X และตัวแปรตาม Y ที่ได้จากขั้นตอนที่ 1 ในหัวข้อ 3.3.1 มาหาพารามิเตอร์ปรับแต่ง (Tunning Parameter) ด้วยวิธี K-fold Cross Validation โดยกำหนดให้ $K = 5$ และใช้ค่าความคลาดเคลื่อนกำลังสองเฉลี่ย (Mean Squared Error: MSE) เป็นตัววัดประสิทธิภาพของพารามิเตอร์ปรับแต่ง โดยการวิเคราะห์ด้วยวิธี L0L2Learn จะต้องหา Tunning Parameter ทั้งหมด 2 ตัวด้วยกันคือ λ_0 และ λ_2 ซึ่งมีขั้นตอนในการหา ดังนี้

1.1) แบ่งข้อมูลออกเป็น 5 ส่วนเท่า ๆ กัน จะได้ T_1, T_2, T_3, T_4, T_5 ให้ข้อมูล 1 ส่วนเป็นชุดข้อมูลทดสอบย่อย และข้อมูลอีก 4 ชุดเป็นข้อมูลฝึกฝนย่อย

1.2) กำหนดค่าพารามิเตอร์ปรับแต่ง $\lambda_2^i > 0$ เมื่อ λ_2^i เป็นค่าคงที่ใด ๆ ที่อยู่ในช่วง [0.0001, 10] ซึ่งจะมีจำนวนทั้งหมด 10 ค่า และกำหนดค่าพารามิเตอร์ปรับแต่ง $\lambda_0^j > 0$ เมื่อ λ_0^j เป็นค่าคงที่ใด ๆ ขึ้นอยู่กับลักษณะของข้อมูล ซึ่งจะมีจำนวนทั้งหมด 100 ค่า

1.3) ทำการวิเคราะห์หาตัวแบบเมื่อ λ_2^i และ λ_0^j เป็น Tunning Parameter และทำการวิเคราะห์หาตัวแบบที่ทำให้ MSE มีค่าต่ำที่สุด จากนั้นนำตัวแบบที่ได้ไปใช้ในการพยากรณ์ค่าตัวแปรตาม ในชุดข้อมูลทดสอบย่อยแล้วคำนวณค่า MSE ของแต่ละชุด ทำเช่นนี้จนครบ 5 ชุดจะได้ค่า MSE ทั้งหมด 5 ค่า จากนั้นหาค่าเฉลี่ยของค่า MSE ที่จุดใด ๆ และทำเช่นนี้ไปเรื่อย ๆ จนครบทุก λ_2^i และ λ_0^j

1.4) เลือก λ_2^i และ λ_0^j ที่ทำให้ค่าเฉลี่ยของค่า MSE มีค่าต่ำที่สุด จะได้ว่าค่า λ_2^i และ λ_0^j นั้นคือพารามิเตอร์ปรับแต่งที่เหมาะสมที่สุด

2) นำค่าพารามิเตอร์ปรับแต่งที่ได้จากข้อ 1 มาหาค่าสัมประสิทธิ์การถดถอยจากข้อมูลทั้งหมดอีกครั้ง จะได้ตัวแบบที่เหมาะสมที่สุด

3) จากนั้นนำตัวแบบที่เหมาะสมที่สุดที่ฝึกฝนด้วยข้อมูลชุดที่ 1 ที่ได้จากขั้นตอนที่ 2 มาคำนวณค่า MSE จากชุดทดสอบ โดยกำหนดให้ข้อมูลชุดที่ 2 ที่ถูกจำลองด้วยเงื่อนไขเดียวกันเป็นข้อมูลชุดทดสอบเพื่อหาค่า MSE และเมื่อใช้ข้อมูลชุดที่ 2 ซึ่งเป็นข้อมูลชุดฝึกฝนในการหาค่า MSE จะกำหนดให้ข้อมูลชุดที่ 3 ที่ถูกจำลองด้วยเงื่อนไขเดียวกันเป็นข้อมูลชุดทดสอบเพื่อหาค่า MSE ทำเช่นนี้ไปเรื่อย ๆ จนตัวแบบที่ได้จากการฝึกฝนด้วยข้อมูลชุดที่ 100 จะใช้ข้อมูลชุดที่ 1 เป็นข้อมูลชุดทดสอบในการหาค่า MSE

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

4) นับจำนวนสัมประสิทธิ์การถดถอยที่เป็น True Positive, False Negative, False Positive, และ True Negative จากนั้นคำนวณหาค่า Precision, Recall, และ AUC ตามลำดับ

5) ทำซ้ำขั้นตอนที่ 1 ถึง 4 จำนวน 100 ครั้งตามการจำลองข้อมูล 100 รอบในแต่ละสถานการณ์

6) หาค่าเฉลี่ยของค่า MSE, Precision, Recall, และ AUC จากผลการทดลองทั้ง 100 ครั้ง

3.3.4 ขั้นตอนการหาตัวประมาณด้วยวิธี L1

1) นำชุดข้อมูลตัวแปรอิสระ X และตัวแปรตาม Y ที่ได้จากขั้นตอนที่ 1 ในหัวข้อ 3.3.1 มาหาพารามิเตอร์ปรับแต่ง (Tunning Parameter) ด้วยวิธี K-fold Cross Validation โดยกำหนดให้ $K = 5$ และใช้ค่าความคลาดเคลื่อนกำลังสองเฉลี่ย (Mean Squared Error: MSE) เป็นตัววัดประสิทธิภาพของพารามิเตอร์ปรับแต่ง โดยมีขั้นตอนในการหา Tunning Parameter คล้ายคลึงกับการหา Tunning Parameter ของวิธี L0Learn คือ

1.1) แบ่งข้อมูลออกเป็น 5 ส่วนเท่า ๆ กัน จะได้ T_1, T_2, T_3, T_4, T_5 ให้ข้อมูล 1 ส่วนเป็นชุดข้อมูลทดสอบย่อย และข้อมูลอีก 4 ชุดเป็นข้อมูลฝึกฝนย่อย

1.2) กำหนดค่าพารามิเตอร์ปรับแต่ง $\lambda_1^i \geq 0$ เมื่อ λ_1^i เป็นค่าคงที่ใด ๆ จากนั้น ณ จุด λ_1^i ใด ๆ ทำการวิเคราะห์หาตัวแบบที่ทำให้ MSE มีค่าต่ำที่สุด

1.3) จากนั้นนำตัวแบบที่ได้จากขั้นตอนที่ 1.2 ไปใช้ในการพยากรณ์ค่าตัวแปรตาม Y ในชุดข้อมูลทดสอบย่อยแล้วคำนวณค่า MSE ของแต่ละชุด ทำเช่นนี้จนครบ 5 ชุดจะได้ค่า MSE ทั้งหมด 5 ค่า จากนั้นหาค่าเฉลี่ยของค่า MSE ที่จุด λ_1^i ใด ๆ

1.4) เลือก t ที่ทำให้ค่าเฉลี่ยของค่า MSE มีค่าต่ำที่สุด จะได้ว่าค่า λ_1^i นั้นคือพารามิเตอร์ปรับแต่งที่เหมาะสมที่สุด

2) นำค่าพารามิเตอร์ปรับแต่งที่เหมาะสมที่สุดมาหาค่าสัมประสิทธิ์การถดถอยจากข้อมูลทั้งหมดอีกครั้ง

3) จากนั้นนำตัวแบบที่เหมาะสมที่สุดที่ฝึกฝนด้วยข้อมูลชุดที่ 1 ที่ได้จากขั้นตอนที่ 2 มาคำนวณค่า MSE จากชุดทดสอบ โดยกำหนดให้ข้อมูลชุดที่ 2 ที่ถูกจำลองด้วยเงื่อนไขเดียวกันเป็นข้อมูลชุดทดสอบเพื่อหาค่า MSE และเมื่อใช้ข้อมูลชุดที่ 2 ซึ่งเป็นข้อมูลชุดฝึกฝนในการหาค่า MSE จะกำหนดให้ข้อมูลชุดที่ 3 ที่ถูกจำลองด้วยเงื่อนไขเดียวกันเป็นข้อมูลชุดทดสอบเพื่อหาค่า MSE ทำเช่นนี้ไปเรื่อย ๆ จนตัวแบบที่ได้จากการฝึกฝนด้วยข้อมูลชุดที่ 100 จะใช้ข้อมูลชุดที่ 1 เป็นข้อมูลชุดทดสอบในการหาค่า MSE

4) นับจำนวนสัมประสิทธิ์การถดถอยที่เป็น True Positive, False Negative, False Positive, และ True Negative จากนั้นคำนวณหาค่า Precision, Recall, และ AUC ตามลำดับ

5) ทำซ้ำขั้นตอนที่ 1 ถึง 4 จำนวน 100 ครั้งตามการจำลองข้อมูล 100 รอบในแต่ละสถานการณ์

6) หาค่าเฉลี่ยของค่า MSE, Precision, Recall, และ AUC จากผลการทดลองทั้ง 100 ครั้ง

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

3.3.5 ขั้นตอนการหาตัวประมาณด้วยวิธี A-L1

1) หาเวกเตอร์ถ่วงน้ำหนักที่เหมาะสมจากวิธี Ridge regression ดังนี้

1.1) แบ่งข้อมูลชุดที่ 1 ออกเป็น 5 ส่วนเท่า ๆ กัน จะได้ T_1, T_2, T_3, T_4, T_5 ให้ข้อมูล 1 ส่วนเป็นชุดข้อมูลทดสอบย่อย และข้อมูลอีก 4 ชุดเป็นข้อมูลฝึกฝนย่อย

1.2) กำหนดค่าพารามิเตอร์ปรับแต่ง $\lambda_2^i \geq 0$ เมื่อ λ_2^i เป็นค่าคงที่ใด ๆ จากนั้น ณ จุด λ_2^i ใด ๆ

1.3) ทำการวิเคราะห์หาตัวแบบจากข้อมูลฝึกฝนย่อยที่ทำให้ MSE มีค่าต่ำที่สุด จากนั้นนำตัวแบบที่ได้ไปใช้ในการพยากรณ์ค่าตัวแปรตาม ในชุดข้อมูลทดสอบย่อยแล้วคำนวณค่า MSE ของแต่ละชุด ทำเช่นนี้จนครบ 5 ชุดจะได้ค่า MSE ทั้งหมด 5 ค่า

1.4) จากนั้นหาค่าเฉลี่ยของค่า MSE ที่จุด λ_2^i ใด ๆ

1.5) เลือก λ_2^i ที่ทำให้ค่าเฉลี่ยของค่า MSE มีค่าต่ำที่สุด จะได้ว่าค่า λ_2^i นั้นคือพารามิเตอร์ปรับแต่งที่เหมาะสมที่สุด

1.6) นำ λ_2^i ที่เหมาะสมที่สุดมาหาค่าสัมประสิทธิ์การถดถอยจากข้อมูลทั้งหมดอีกครั้ง ได้เวกเตอร์ถ่วงน้ำหนักที่เหมาะสม $\hat{w}_j = \frac{1}{\left|\hat{\beta}_j^{Ridge}\right|^{\gamma}}$

2) หาพารามิเตอร์ปรับแต่ง (Tunning Parameter) λ_1 เพื่อใช้ในการวิเคราะห์ตัวแบบ โดยมีขั้นตอนดังนี้

2.1) แบ่งข้อมูลออกเป็น 5 ส่วนเท่า ๆ กัน จะได้ T_1, T_2, T_3, T_4, T_5 ให้ข้อมูล 1 ส่วนเป็นชุดข้อมูลทดสอบย่อย และข้อมูลอีก 4 ชุดเป็นข้อมูลฝึกฝนย่อย

2.2) นำเวกเตอร์ถ่วงน้ำหนักที่ได้จากข้อ 1.6) มาใช้ในการประมาณค่าสัมประสิทธิ์การถดถอยด้วยวิธี L1 โดยกำหนดค่าพารามิเตอร์ปรับแต่ง $\lambda_1^i \geq 0$ เมื่อ λ_1^i เป็นค่าคงที่ใด ๆ จากนั้น ณ จุด λ_1^i ใด ๆ ทำการวิเคราะห์หาตัวแบบที่ทำให้ MSE มีค่าต่ำที่สุด โดยเวกเตอร์ถ่วงน้ำหนักที่ใช้ในการหาค่า λ_1^i ที่เหมาะสมนั้นจะไม่เปลี่ยนแปลงในการทำ Cross-validation ในแต่ละรอบ

2.3) จากนั้นนำตัวแบบที่ได้จากขั้นตอนที่ 1.2 ไปใช้ในการพยากรณ์ค่าตัวแปรตาม Y ในชุดข้อมูลทดสอบย่อยแล้วคำนวณค่า MSE ของแต่ละชุด ทำเช่นนี้จนครบ 5 ชุดจะได้ค่า MSE ทั้งหมด 5 ค่า จากนั้นหาค่าเฉลี่ยของค่า MSE ที่จุด λ_1^i ใด ๆ

2.4) เลือก t ที่ทำให้ค่าเฉลี่ยของค่า MSE มีค่าต่ำที่สุด จะได้ว่าค่า λ_1^i นั้นคือพารามิเตอร์ปรับแต่งที่เหมาะสมที่สุด

2.5) นำค่าพารามิเตอร์ปรับแต่งที่เหมาะสมที่สุดมาหาค่าสัมประสิทธิ์การถดถอยจากข้อมูลทั้งหมดอีกครั้ง จะได้ตัวแบบที่เหมาะสมที่สุดสำหรับข้อมูลชุดนั้น ๆ

3) จากนั้นนำตัวแบบที่เหมาะสมที่สุดที่ฝึกฝนด้วยข้อมูลชุดที่ 1 ที่ได้จากขั้นตอนที่ 2 มาคำนวณค่า MSE จากชุดทดสอบ โดยกำหนดให้ข้อมูลชุดที่ 2 ที่ถูกจำลองด้วยเงื่อนไขเดียวกันเป็นข้อมูลชุดทดสอบเพื่อหาค่า MSE และเมื่อใช้ข้อมูลชุดที่ 2 ซึ่งเป็นข้อมูลชุดฝึกฝนในการหาค่า MSE จะกำหนดให้ข้อมูลชุดที่ 3 ที่ถูกจำลองด้วยเงื่อนไขเดียวกันเป็นข้อมูลชุดทดสอบเพื่อหาค่า MSE ทำเช่นนี้ไปเรื่อย ๆ จนตัวแบบที่ได้จากการฝึกฝนด้วยข้อมูลชุดที่ 100 จะใช้ข้อมูลชุดที่ 1 เป็นข้อมูลชุดทดสอบในการหาค่า MSE

4) นับจำนวนสัมประสิทธิ์การถดถอยที่เป็น True Positive, False Negative, False Positive, และ True Negative จากนั้นคำนวณหาค่า Precision, Recall, และ AUC ตามลำดับ

5) ทำซ้ำขั้นตอนที่ 1 ถึง 4 จำนวน 100 ครั้งตามการจำลองข้อมูล 100 รอบในแต่ละสถานการณ์

6) หาค่าเฉลี่ยของค่า MSE, Precision, Recall, และ AUC จากผลการทดลองทั้ง 100 ครั้ง

3.3.6 วิธีการดำเนินงานวิจัยด้วยวิธี A-L1L2

1) หาเวกเตอร์ถ่วงน้ำหนักที่เหมาะสมจากวิธี Ridge regression ดังนี้

1.1) แบ่งข้อมูลชุดที่ 1 ออกเป็น 5 ส่วนเท่า ๆ กัน จะได้ T_1, T_2, T_3, T_4, T_5 ให้ข้อมูล 1 ส่วนเป็นชุดข้อมูลทดสอบย่อย และข้อมูลอีก 4 ชุดเป็นข้อมูลฝึกฝนย่อย

1.2) กำหนดค่าพารามิเตอร์ปรับแต่ง $\lambda_2^i \geq 0$ เมื่อ λ_2^i เป็นค่าคงที่ใด ๆ จากนั้น ณ จุด λ_2^i ใด ๆ

CHULALONGKORN UNIVERSITY

1.3) ทำการวิเคราะห์หาตัวแบบจากข้อมูลฝึกฝนย่อยที่ทำให้ **MSE** มีค่าต่ำที่สุด จากนั้นนำตัวแบบที่ได้ไปใช้ในการพยากรณ์ค่าตัวแปรตาม ในชุดข้อมูลทดสอบย่อยแล้วคำนวณค่า **MSE** ของแต่ละชุด ทำเช่นนี้จนครบ 5 ชุดจะได้ค่า **MSE** ทั้งหมด 5 ค่า

1.4) จากนั้นหาค่าเฉลี่ยของค่า **MSE** ที่จุด λ_2^i ใด ๆ

1.5) เลือก λ_2^i ที่ทำให้ค่าเฉลี่ยของค่า **MSE** มีค่าต่ำที่สุด จะได้ว่าค่า λ_2^i นั้นคือพารามิเตอร์ปรับแต่งที่เหมาะสมที่สุด

1.6) นำ λ_2^i ที่เหมาะสมที่สุดมาหาค่าสัมประสิทธิ์การถดถอยจากข้อมูลทั้งหมดอีกครั้ง ได้เวกเตอร์ถ่วงน้ำหนักที่เหมาะสม $\hat{w}_j = \frac{1}{\left|\hat{\beta}_j^{Ridge}\right|^{\gamma}}$

2) หาพารามิเตอร์ปรับแต่ง (Tunning Parameter) λ_1 และ λ_2 เพื่อใช้ในการวิเคราะห์ตัวแบบ โดยมีขั้นตอนดังนี้

2.1) กำหนดค่าพารามิเตอร์ปรับแต่ง $\lambda_2^i > 0$ เมื่อ λ_2^i เป็นค่าคงที่ใด ๆ ที่อยู่ในช่วง [0.1, 0.9] ซึ่งจะมีจำนวนทั้งหมด 10 ค่า และกำหนดค่าพารามิเตอร์ปรับแต่ง $\lambda_1^j > 0$ เมื่อ λ_1^j เป็นค่าคงที่ใด ๆ จะมีจำนวนทั้งหมด 100 ค่า

2.2) นำเวกเตอร์ถ่วงน้ำหนักที่ได้จากข้อ 1.6) มาใช้ในการวิเคราะห์หาตัวแบบเมื่อ λ_2^i และ λ_1^j เป็น Tunning Parameter และทำการวิเคราะห์หาตัวแบบที่ทำให้ MSE มีค่าต่ำที่สุด จากนั้นนำตัวแบบที่ได้ไปใช้ในการพยากรณ์ค่าตัวแปรตาม ในชุดข้อมูลทดสอบย่อยแล้วคำนวณค่า MSE ของแต่ละชุด ทำเช่นนี้จนครบ 5 ชุดจะได้ค่า MSE ทั้งหมด 5 ค่า จากนั้นหาค่าเฉลี่ยของค่า MSE ที่จุดใด ๆ และทำเช่นนี้ไปเรื่อย ๆ จนครบทุก λ_2^i และ λ_1^j โดยเวกเตอร์ถ่วงน้ำหนักที่ใช้ในการหาค่า λ_2^i และ λ_1^j ที่เหมาะสมนั้นจะไม่เปลี่ยนแปลงในการทำ Cross-validation ในแต่ละรอบ

2.3) เลือก λ_2^i และ λ_1^j ที่ทำให้ค่าเฉลี่ยของค่า MSE มีค่าต่ำที่สุด จะได้ว่าค่า λ_2^i และ λ_1^j นั้นคือพารามิเตอร์ปรับแต่งที่เหมาะสมที่สุด

2.4) นำค่าพารามิเตอร์ปรับแต่งที่เหมาะสมที่สุดมาหาค่าสัมประสิทธิ์การถดถอยจากข้อมูลทั้งหมดอีกครั้ง จะได้ตัวแบบที่เหมาะสมที่สุดสำหรับข้อมูลชุดนั้น ๆ

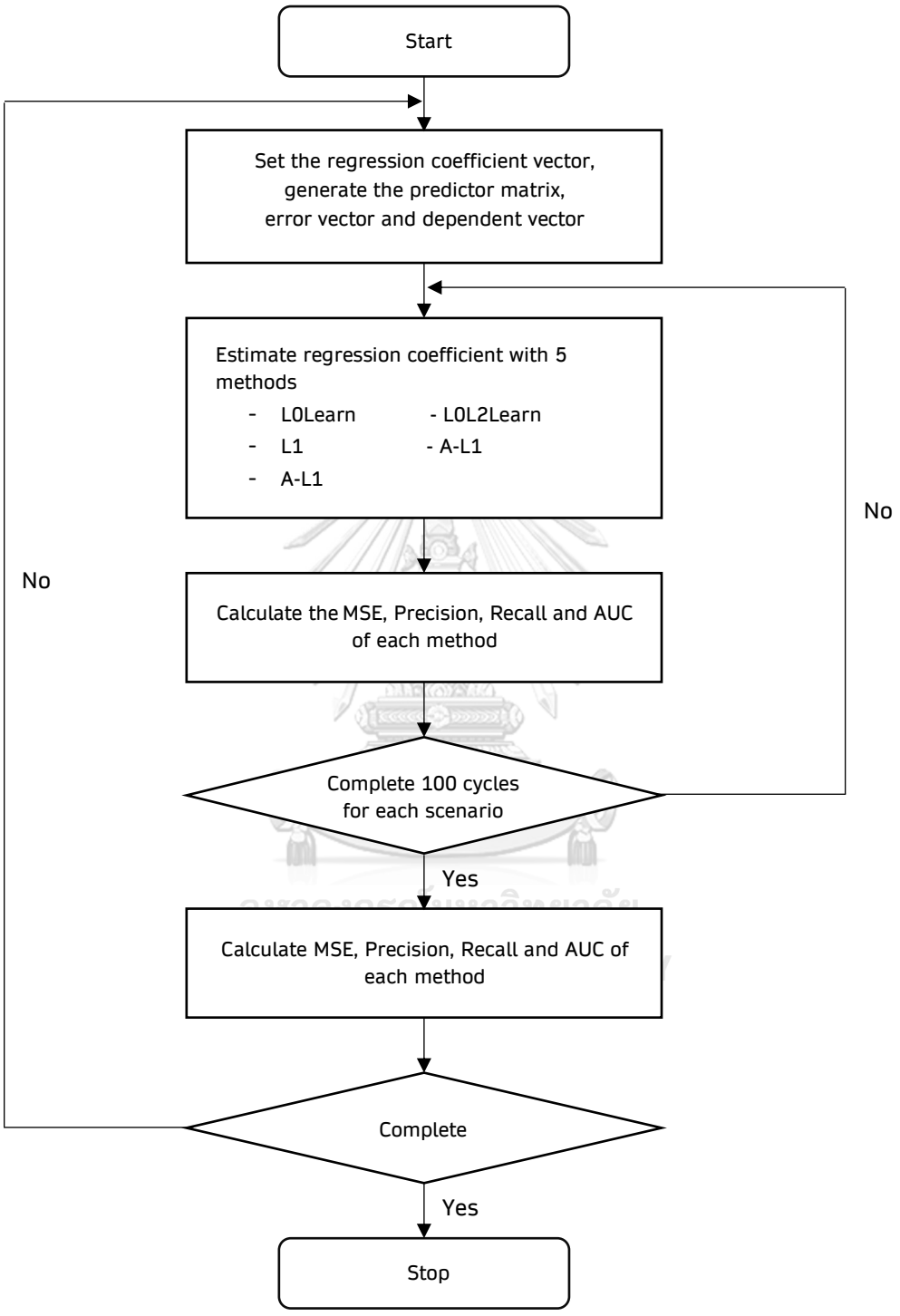
3) จากนั้นนำตัวแบบที่เหมาะสมที่สุดที่ฝึกฝนด้วยข้อมูลชุดที่ 1 ที่ได้จากขั้นตอนที่ 2 มาคำนวณค่า MSE จากชุดทดสอบ โดยกำหนดให้ข้อมูลชุดที่ 2 ที่ถูกจำลองด้วยเงื่อนไขเดียวกันเป็นข้อมูลชุดทดสอบเพื่อหาค่า MSE และเมื่อใช้ข้อมูลชุดที่ 2 ซึ่งเป็นข้อมูลชุดฝึกฝนในการหาค่า MSE จะกำหนดให้ข้อมูลชุดที่ 3 ที่ถูกจำลองด้วยเงื่อนไขเดียวกันเป็นข้อมูลชุดทดสอบเพื่อหาค่า MSE ทำเช่นนี้ไปเรื่อย ๆ จนตัวแบบที่ได้จากการฝึกฝนด้วยข้อมูลชุดที่ 100 จะใช้ข้อมูลชุดที่ 1 เป็นข้อมูลชุดทดสอบในการหาค่า MSE

4) นับจำนวนสัมประสิทธิ์การถดถอยที่เป็น True Positive, False Negative, False Positive, และ True Negative จากนั้นคำนวณหาค่า Precision, Recall, และ AUC ตามลำดับ

5) ทำซ้ำขั้นตอนที่ 1 ถึง 4 จำนวน 100 ครั้งตามการจำลองข้อมูล 100 รอบในแต่ละสถานการณ์

6) หาค่าเฉลี่ยของค่า MSE, Precision, Recall, และ AUC จากผลการทดลองทั้ง 100 ครั้ง

3.4 แผนภาพการทำงานของโปรแกรม



รูปที่ 3.2 แผนภาพการทำงานของโปรแกรม

บทที่ 4
ผลการวิจัย

การวิจัยครั้งนี้เป็นการวิจัยเชิงทดลอง มีวัตถุประสงค์เพื่อศึกษาพฤติกรรม และเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าสัมประสิทธิ์การถดถอยเชิงเส้นสำหรับข้อมูลที่มีมิติสูง ด้วยวิธี Penalized Regression วิธี L0Learn และวิธี L0L2Learn เปรียบเทียบกับวิธี Penalized Regression ที่ได้รับความนิยมในปัจจุบัน ได้แก่ วิธี L1 วิธี A-L1 และ วิธี A-L1L2

ในการเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าสัมประสิทธิ์การถดถอยเชิงเส้นสำหรับข้อมูลที่มีมิติสูง จะทำการเปรียบเทียบผลลัพธ์โดยมุ่งเน้นไปที่ 2 กรณีคือ 1 เมื่อข้อมูลมีตัวแปรอิสระมีความสัมพันธ์กันแบบยกกำลัง (Exponential Correlation) เมื่อ ρ = 0.0, 0.5 และ 0.9 โดยกำหนดให้ค่า SNR คงที่ และในกรณีที่ 2 เมื่อข้อมูลมีสัญญาณรบกวน (Noise) แตกต่างกัน โดยกำหนดให้ SNR = 0.1, 0.5, 1, 5, 10 และ 20 และกำหนดให้ค่า ρ คงที่ โดยแต่ละกรณีจะทำการจำลองข้อมูลที่มีขนาดตัวอย่างเท่ากับ 100 ตัวอย่าง และจำนวนตัวแปรอิสระเท่ากับ 1000 ตัว

สำหรับการเปรียบเทียบประสิทธิภาพจะเปรียบเทียบใน 2 ด้านเป็นหลัก คือ 1) เปรียบเทียบประสิทธิภาพด้านการพยากรณ์ ซึ่งพิจารณาจากความถูกต้องของการประมาณค่าสัมประสิทธิ์การถดถอยซึ่งวัดจากค่าเฉลี่ยของค่าคลาดเคลื่อนการทำนาย (AMSE) และ 2) ความถูกต้องในการคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบ ซึ่งพิจารณาจากค่าเฉลี่ยของค่า Precision ค่าเฉลี่ยของค่า Recall และค่าเฉลี่ยของค่า AUC ซึ่งได้จากการจำลองข้อมูลซ้ำจำนวน 100 รอบในแต่ละสถานการณ์ ซึ่งผลการวิจัยในแต่ละกรณีมีรายละเอียด ดังนี้

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

4.1 ประสิทธิภาพด้านการพยากรณ์

การเปรียบเทียบประสิทธิภาพด้านการพยากรณ์ของทั้ง 5 วิธี เมื่อข้อมูลมีระดับ SNR และตัวแปรอิสระไม่มีความสัมพันธ์กันในระดับต่าง ๆ มีรายละเอียดของผลการวิเคราะห์ ดังนี้

1) เมื่อข้อมูลมีระดับ SNR แตกต่างกัน และตัวแปรอิสระไม่มีความสัมพันธ์กัน (ρ = 0.0)

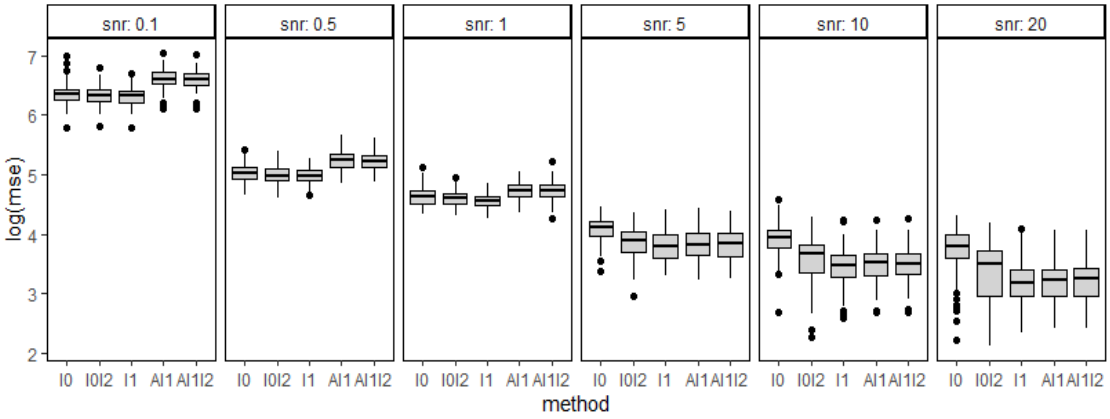
พิจารณาผลการวิเคราะห์จากข้อมูลชุดทดสอบจากตารางที่ 4.1 แสดงค่าเฉลี่ยความคลาดเคลื่อนกำลังสอง (AMSE) และรูปที่ 4.1 พบว่าวิธี L1 มีค่า AMSE ต่ำที่สุด (ค่า AMSE ยิ่งน้อยยิ่งดี) หรือกล่าวได้ว่ามีประสิทธิภาพในการพยากรณ์มากที่สุดแม้ว่าข้อมูลจะมีค่า SNR เปลี่ยนแปลงไป ส่วนวิธี L0Learn และวิธี L0L2Learn นั้นเมื่อข้อมูลมีค่า SNR น้อยกว่าหรือเท่ากับ 1 ทั้งสองวิธีจะมีประสิทธิภาพในการพยากรณ์ใกล้เคียงกันและมีประสิทธิภาพสูงกว่าตัวแบบที่ได้จากวิธี A-L1 และวิธี A-L1L2 แต่เมื่อข้อมูลมีค่า SNR เพิ่มสูงขึ้นวิธี L0L2Learn จะมีประสิทธิภาพในการพยากรณ์มากกว่า

วิธี L0Learn อย่างไรก็ตามทั้งวิธี L0Learn และวิธี L0L2Learn จะมีค่า AMSE มากกว่าวิธี A-L1 และวิธี A-L1L2 นั่นคือ เมื่อข้อมูลมีสัญญาณรบกวนน้อยวิธี L0Learn และวิธี L0L2Learn จะมีประสิทธิภาพต่ำกว่าวิธี A-L1 และวิธี A-L1L2 โดยสรุปแล้ว ในกรณีที่ตัวแปรอิสระไม่มีความสัมพันธ์กันวิธี L1 มีประสิทธิภาพในการพยากรณ์สูงที่สุดในทุกกรณี ส่วนวิธี L0Learn และวิธี L0L2Learn จะมีประสิทธิภาพในการพยากรณ์เมื่อข้อมูลมีสัญญาณรบกวนสูงหรือมีค่า SNR ต่ำ ส่วนวิธี A-L1 และวิธี A-L1L2 จะมีประสิทธิภาพสูงขึ้นเมื่อข้อมูลมีสัญญาณรบกวนต่ำหรือมีค่า SNR สูง

ตารางที่ 4.1 ค่าเฉลี่ยของค่าความคลาดเคลื่อนกำลังสอง (AMSE) จาก 100 ชุดข้อมูล เมื่อตัวแปรอิสระไม่มีความสัมพันธ์กัน

วิธี	ค่าสถิติ	ค่า SNR					
		0.1	0.5	1	5	10	20
n = 100, ρ = 0.0							
L0Learn	Mean	581.558	152.610	103.236	59.360	50.799	44.238
	SD.	108.868	24.359	15.986	10.799	12.551	14.084
L0L2Learn	Mean	567.274	149.300	100.987	48.434	37.505	31.310
	SD.	88.299	24.242	13.858	12.027	12.012	12.810
L1	Mean	556.639	146.438	96.412	45.469	33.686	25.755
	SD.	82.686	19.996	11.617	11.673	10.631	10.428
A-L1	Mean	750.546	189.098	114.517	47.106	34.074	26.089
	SD.	120.357	31.213	16.861	11.538	9.954	8.799
A-L1L2	Mean	745.291	186.677	113.871	47.086	34.063	26.370
	SD.	115.160	28.126	18.652	11.449	10.110	8.980

CHULALONGKORN UNIVERSITY



รูปที่ 4.1 Boxplot แสดงค่าความคลาดเคลื่อนกำลังสอง (MSE) เมื่อตัวแปรอิสระไม่มีความสัมพันธ์กัน

2) เมื่อข้อมูลมีระดับ SNR แตกต่างกัน และตัวแปรอิสระมีความสัมพันธ์กันปานกลาง (ρ = 0.5)

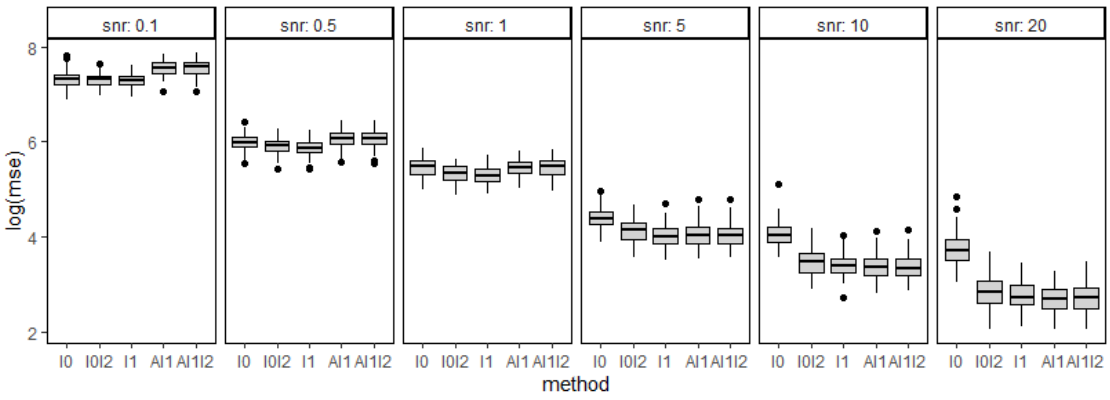
พิจารณาผลการวิเคราะห์จากข้อมูลชุดทดสอบจากตารางที่ 4.2 แสดงค่าเฉลี่ยความคลาดเคลื่อนกำลังสอง (AMSE) และรูปที่ 4.2 พบว่าเมื่อข้อมูลมีค่า SNR น้อยกว่าเท่ากับ 5 วิธี L1 จะมีค่า AMSE ต่ำที่สุดหรือมีประสิทธิภาพในการพยากรณ์มากที่สุด แต่เมื่อค่า SNR เพิ่มสูงขึ้นวิธี A-L1 และวิธี A-L1L2 จะมีค่า AMSE น้อยกว่าวิธีอื่น ๆ กล่าวได้ว่าในกรณีที่ข้อมูลที่ค่า SNR มากกว่าเท่ากับ 10 แล้ว วิธี A-L1 และวิธี A-L1L2 จะมีประสิทธิภาพในการพยากรณ์สูงที่สุด นอกจากนี้สังเกตได้ว่าค่า AMSE ที่ได้จากวิธี A-L1 และวิธี A-L1L2 จะมีค่าใกล้เคียงกันมากในทุกกรณี ส่วนวิธี L0L2Learn นั้นพบว่าจะมีประสิทธิภาพในการพยากรณ์สูงเป็นลำดับที่สองรองจากวิธี L1 เมื่อข้อมูลมีค่า SNR ต่ำหรือมีสัญญาณรบกวนสูง ส่วนผลการวิเคราะห์ด้วยตัวแบบที่ได้จากวิธี L0Learn นั้นพบว่าเมื่อข้อมูลมีสัญญาณรบกวนต่ำค่า AMSE จะมีค่าสูงกว่าวิธี L0L2Learn เพียงเล็กน้อย และมีประสิทธิภาพในการพยากรณ์สูงเป็นลำดับ 3 แต่เมื่อข้อมูลมีค่า SNR เพิ่มมากขึ้นวิธี L0Learn นั้นจะมีความสามารถในการพยากรณ์น้อยที่สุดเมื่อเปรียบเทียบกับวิธีอื่น ๆ

นอกจากนี้เมื่อเปรียบเทียบผลการวิเคราะห์ในกรณีนี้ซึ่งตัวแปรอิสระมีความสัมพันธ์กันปานกลางกับกรณีที่ตัวแปรอิสระไม่มีความสัมพันธ์กัน พบว่าทั้งสองกรณีมีความคล้ายคลึงกัน คือ วิธี L0L2Learn และวิธี L0Learn จะมีประค่า AMSE ใกล้เคียงกันเมื่อข้อมูลมีค่า SNR น้อยและมีประสิทธิภาพในการพยากรณ์เป็นลำดับที่ 2 และ 3 ตามลำดับ แต่เมื่อข้อมูลมีค่า SNR เพิ่มสูงขึ้นวิธี L0L2Learn จะมีค่า AMSE น้อยกว่าวิธี L0Learn อย่างชัดเจน อย่างไรก็ตามทั้งวิธี L0L2Learn และวิธี L0Learn ยังมีประสิทธิภาพในการพยากรณ์น้อยกว่าวิธีที่อยู่ในกลุ่มที่ใช้ L_1norm ซึ่งได้แก่วิธี L1 วิธี A-L1 และวิธี A-L1L2 ส่วนความแตกต่างของผลการวิเคราะห์กรณีตัวแปรอิสระมีความสัมพันธ์กันปานกลางกับกรณีที่ตัวแปรอิสระไม่มีความสัมพันธ์กัน คือ ในกรณีที่ตัวแปรอิสระไม่มีความสัมพันธ์กันวิธี L1 จะมีประสิทธิภาพในการพยากรณ์สูงที่สุดในทุกระดับ SNR แต่เมื่อตัวแปรอิสระมีความสัมพันธ์กันปานกลางและมีค่า SNR มากกว่าหรือเท่ากับ 10 แล้ววิธี A-L1 และวิธี A-L1L2 จะมีประสิทธิภาพในการพยากรณ์สูงที่สุด

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

ตารางที่ 4.2 ค่าเฉลี่ยของค่าความคลาดเคลื่อนกำลังสอง (AMSE) จาก 100 ชุดข้อมูล เมื่อตัวแปรอิสระมีความสัมพันธ์กันปานกลาง

วิธี	ค่าสถิติ	ค่า SNR					
		0.1	0.5	1	5	10	20
n = 100, ρ = 0.5							
L0Learn	Mean	1,534.045	402.177	238.661	84.712	59.560	44.347
	SD.	267.992	66.194	46.768	18.799	17.335	16.008
L0L2Learn	Mean	1,500.266	366.863	211.037	64.546	33.859	18.328
	SD.	218.630	62.900	36.661	16.644	10.442	6.070
L1	Mean	1,475.635	360.574	201.882	56.083	31.049	16.424
	SD.	208.574	64.028	35.860	12.770	8.133	4.664
A-L1	Mean	1,945.275	434.210	237.995	59.159	30.243	15.714
	SD.	280.424	84.644	43.522	15.229	7.928	4.465
A-L1L2	Mean	1,941.372	437.631	235.657	58.118	30.254	15.827
	SD.	295.601	86.545	43.462	14.721	8.396	4.590



รูปที่ 4.2 Boxplot แสดงค่าความคลาดเคลื่อนกำลังสอง (MSE) เมื่อตัวแปรอิสระมีความสัมพันธ์กันปานกลาง

3) เมื่อข้อมูลมีระดับ SNR แตกต่างกัน และตัวแปรอิสระมีความสัมพันธ์กันสูง (ρ = 0.9)

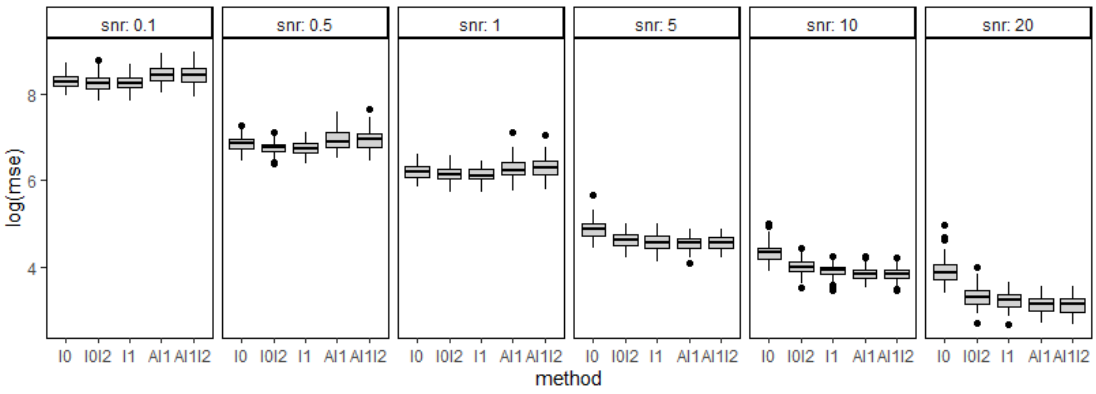
พิจารณาผลการวิเคราะห์จากข้อมูลชุดทดสอบจากตารางที่ 4.3 แสดงค่าเฉลี่ยความคลาดเคลื่อนกำลังสอง (AMSE) และรูปที่ 4.3 พบว่าเมื่อข้อมูลมีค่า SNR น้อยกว่าเท่ากับ 1 วิธี L1 จะมีค่า AMSE ต่ำที่สุดหรือมีประสิทธิภาพในการพยากรณ์มากที่สุด แต่เมื่อค่า SNR เพิ่มสูงขึ้นและมีค่ามากกว่าหรือเท่ากับ 5 วิธี A-L1 และวิธี A-L1L2 จะมีค่า AMSE น้อยกว่าวิธีอื่น ๆ กล่าวได้ว่ามี

ประสิทธิภาพในการพยากรณ์สูงที่สุด นอกจากนี้สังเกตได้ว่าค่า AMSE ที่ได้จากวิธี A-L1 และวิธี A-L1L2 จะมีค่าใกล้เคียงกันมากในทุกกรณี ส่วนวิธี L0L2Learn นั้นพบว่าจะมีประสิทธิภาพในการพยากรณ์สูงเป็นลำดับ 2 รองจากวิธี L1 เมื่อข้อมูลมีค่า SNR ต่ำกว่าหรือเท่ากับ 1 ส่วนผลการวิเคราะห์ด้วยตัวแบบที่ได้จากวิธี L0Learn นั้นพบว่าเมื่อข้อมูลมีสัญญาณรบกวนต่ำ วิธี L0Learn จะมีประสิทธิภาพในการพยากรณ์สูงเป็นลำดับ 3 รองจากวิธี L1 และวิธี L0L2Learn แต่เมื่อข้อมูลมีค่า SNR เพิ่มมากขึ้นวิธี L0Learn นั้นจะมีความสามารถในการพยากรณ์น้อยที่สุดเมื่อเปรียบเทียบกับวิธีอื่น ๆ

นอกจากนี้เมื่อเปรียบเทียบผลการวิเคราะห์ของข้อมูลมีตัวแปรอิสระมีระดับความสัมพันธ์แตกต่างกัน พบว่าผลการวิเคราะห์ในกรณีข้อมูลมีตัวแปรอิสระที่มีความสัมพันธ์กันสูงคล้ายกับผลการวิเคราะห์ในกรณีข้อมูลมีตัวแปรอิสระที่มีความสัมพันธ์กันระดับปานกลางเป็นอย่างมาก แตกต่างกันเพียง ในกรณีที่แปรอิสระที่มีความสัมพันธ์กันสูงวิธี A-L1 และวิธี A-L1L2 จะมีประสิทธิภาพในการพยากรณ์สูงเที่สุดเมื่อข้อมูลมีค่า SNR มากกว่าหรือเท่ากับ 5 แต่ในกรณีที่แปรอิสระที่มีความสัมพันธ์กันปานกลางวิธี A-L1 และวิธี A-L1L2 จะมีประสิทธิภาพในการพยากรณ์สูงเที่สุดเมื่อข้อมูลมีค่า SNR มากกว่าหรือเท่ากับ 10

ตารางที่ 4.3 ค่าเฉลี่ยของค่าความคลาดเคลื่อนกำลังสอง (AMSE) จาก 100 ชุดข้อมูล เมื่อตัวแปรอิสระมีความสัมพันธ์กันสูง

วิธี	ค่าสถิติ	ค่า SNR					
		0.1	0.5	1	5	10	20
n = 100, ρ = 0.9							
L0Learn	Mean	4,135.423	962.536	501.567	134.461	78.978	51.587
	SD.	683.883	153.451	91.366	31.371	19.213	17.011
L0L2Learn	Mean	3,979.891	879.455	477.153	102.961	55.374	28.471
	SD.	687.134	124.957	78.733	17.428	9.058	6.862
L1	Mean	3,929.449	866.607	466.494	98.848	50.507	25.812
	SD.	644.459	136.033	72.686	17.126	7.682	4.874
A-L1	Mean	4,802.783	1,055.276	549.380	95.447	46.951	23.096
	SD.	954.498	267.143	130.396	14.407	6.907	3.867
A-L1L2	Mean	4,808.109	1,063.932	554.337	96.225	46.809	23.014
	SD.	1,005.378	260.556	127.182	14.477	6.989	3.916



รูปที่ 4.3 Boxplot แสดงค่าความคลาดเคลื่อนกำลังสอง (MSE) เมื่อตัวแปรอิสระมีความสัมพันธ์กันสูง

4.2 ประสิทธิภาพด้านการคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบ

ในการวัดประสิทธิภาพด้านการคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบ จะใช้ตัววัดทั้งหมด 3 ตัว ด้วยกัน คือ 1) ค่าเฉลี่ยของค่าความแม่นยำ (Average of Precision) 2) ค่าเฉลี่ยของค่าความระลึก (Average of Recall) และ 3) ค่าเฉลี่ยของค่าพื้นที่ใต้กราฟ ROC (Average of AUC Score) ซึ่งคำนวณจากข้อมูล 100 ชุด โดยมีผลในการวิเคราะห์ในแต่ละกรณี ดังนี้

4.2.1 ค่าเฉลี่ยของค่าความแม่นยำ (Average of Precision)

ค่าเฉลี่ยของค่าความแม่นยำ (Average of Precision) เป็นการวัดเป็นการวัดความแม่นยำในการคัดเลือกตัวแปร โดยเฉลี่ยแล้วตัวแบบสามารถคัดเลือกตัวแปรได้ถูกต้องเป็นสัดส่วนเท่าใดเมื่อเปรียบเทียบกับจำนวนตัวแปรที่ถูกคัดเลือกเข้าสู่ตัวแบบทั้งหมด เหมาะกับการวิเคราะห์ที่ต้องการตัวแบบที่ต้องการลดจำนวน False Positive หรือตัวแปรที่ตัวแบบทำนายให้มีค่าสัมประสิทธิ์ไม่เป็นศูนย์แต่ในความเป็นจริงแล้วมีค่าสัมประสิทธิ์เป็นศูนย์ สำหรับผลการวิเคราะห์เมื่อข้อมูลมีระดับ SNR และตัวแปรอิสระไม่มีความสัมพันธ์กันในระดับต่าง ๆ มีรายละเอียด ดังนี้

1) เมื่อข้อมูลมีระดับ SNR แตกต่างกัน และตัวแปรอิสระไม่มีความสัมพันธ์กัน (ρ = 0.0)

พิจารณาผลการวิเคราะห์จากตารางที่ 4.4 ซึ่งแสดงค่าเฉลี่ยของค่า Precision และรูปที่ 4.4 จะเห็นได้ว่าวิธี L0Learn มีค่าเฉลี่ยของค่า Precision สูงที่สุด หรือมีความแม่นยำในการคัดเลือกตัวแปรเข้าสู่ตัวแบบมากที่สุดในเกือบทุกกรณีที่ข้อมูลมีค่า SNR แตกต่างกัน ยกเว้นกรณีที่ข้อมูลมีค่า SNR เท่ากับ 0.5 ส่วนวิธีที่มีค่าเฉลี่ยของค่า Precision โดยเฉลี่ยสูงเป็นลำดับที่ 2 รองจากวิธี

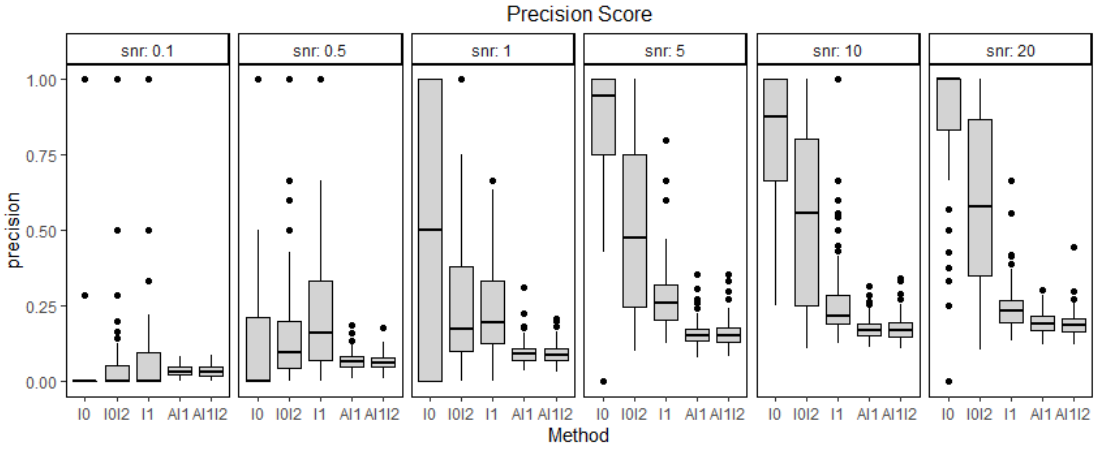
L0Learn คือวิธี L0L2Learn รองลงมาคือวิธี L1 ส่วนวิธี A-L1 และ A-L1L2 นั้นสังเกตเห็นได้อย่างชัดเจนว่ามีประสิทธิภาพใกล้เคียงกัน และมีค่าเฉลี่ยของค่า Precision ค่อนข้างต่ำ

เมื่อค่า SNR มีค่าเพิ่มสูงขึ้น ทุกวิธีจะมีค่า Precision เพิ่มมากขึ้น นั่นคือเมื่อข้อมูลมีสัญญาณรบกวนน้อยความสามารถในการคัดเลือกตัวแปรเข้าสู่ตัวแบบจะเพิ่มสูงขึ้นนั่นเอง แต่เป็นที่น่าสังเกต คือ เมื่อ SNR เพิ่มมากขึ้นค่าเฉลี่ยของค่า Precision ที่ได้จากการวิเคราะห์ด้วยวิธี L0Learn และ L0L2Learn จะเพิ่มสูงขึ้นอย่างรวดเร็ว แตกต่างจากวิธี L1 A-L1 และ A-L1L2 ที่ค่าเฉลี่ยของค่า Precision จะค่อย ๆ เพิ่มขึ้น และมีค่าไม่มากนักแม้ว่าข้อมูลจะมีค่า SNR สูงเท่ากับ 20 ที่เป็นเช่นนี้เพราะวิธี L0Learn และ L0L2Learn จะมีการคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบค่อนข้างน้อย โดยเฉพาะวิธี L0Learn ที่มักมีการคัดเลือกตัวแปรเข้าสู่ตัวแบบน้อยกว่าจำนวนตัวแปรอิสระที่มีค่าสัมประสิทธิ์ไม่เท่ากับศูนย์ที่แท้จริง ทำให้เมื่อคิดเป็นสัดส่วนระหว่างตัวแปรที่ตัวแบบที่ไม่เป็นศูนย์แท้แท้จริงที่ถูกคัดเลือกเข้าสู่ตัวแบบ (True Positive) ส่วนด้วยตัวแปรทั้งหมดที่ถูกคัดเลือกเข้าสู่ตัวแบบทั้งหมด (True Positive + False Positive) ที่มีจำนวนน้อย ทำให้ค่าเฉลี่ยของค่า Precision ที่ค่อนข้างสูง ตรงข้างกับวิธี L1 วิธี A-L1 และวิธี A-L1L2 ที่มักจะคัดเลือกตัวแปรเข้าสู่ตัวแบบเป็นจำนวนมาก ทำให้ค่าเฉลี่ยของค่า Precision มีค่าน้อย

ตารางที่ 4.4 ค่าเฉลี่ยของค่า Precision จาก 100 ชุดข้อมูล เมื่อตัวแปรอิสระไม่มีความสัมพันธ์กัน

วิธี	ค่าสถิติ	ค่า SNR					
		0.1	0.5	1	5	10	20
n = 100, ρ = 0.0							
L0Learn	Mean	0.148	0.195	0.524	0.811	0.800	0.863
	SD.	0.351	0.362	0.413	0.268	0.230	0.239
L0L2Learn	Mean	0.078	0.180	0.297	0.514	0.542	0.598
	SD.	0.201	0.237	0.288	0.294	0.292	0.291
L1	Mean	0.105	0.222	0.242	0.275	0.257	0.243
	SD.	0.223	0.224	0.170	0.111	0.128	0.081
A-L1	Mean	0.035	0.067	0.094	0.162	0.177	0.192
	SD.	0.020	0.028	0.040	0.044	0.038	0.036
A-L1L2	Mean	0.036	0.066	0.095	0.161	0.178	0.190
	SD.	0.021	0.026	0.036	0.046	0.042	0.043

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University



รูปที่ 4.4 Boxplot แสดงค่า Precision เมื่อตัวแปรอิสระไม่มีความสัมพันธ์กัน

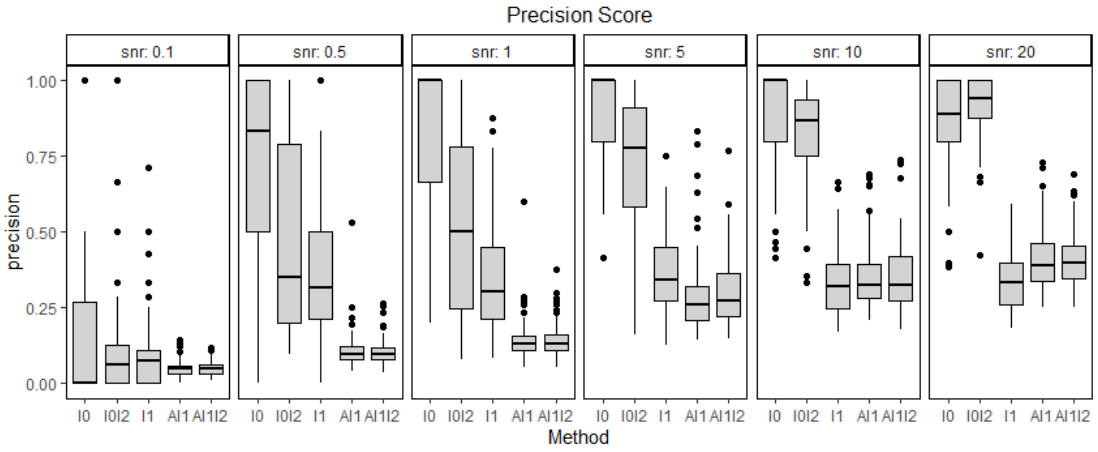
2) เมื่อข้อมูลมีระดับ SNR แตกต่างกัน และตัวแปรอิสระมีความสัมพันธ์กันปานกลาง (ρ = 0.5)

พิจารณาผลการวิเคราะห์จากตารางที่ 4.5 ซึ่งแสดงค่าเฉลี่ยของค่า Precision และรูปที่ 4.5 จะเห็นได้ว่าวิธี L0Learn มีค่าเฉลี่ยของค่า Precision สูงที่สุด หรือมีความแม่นยำในการคัดเลือกตัวแปรเข้าสู่ตัวแบบมากที่สุดในเกือบทุกกรณีที่ข้อมูลมีค่า SNR แตกต่างกัน ยกเว้นกรณีที่ข้อมูลมีค่า SNR เท่ากับ 20 ส่วนวิธีที่มีค่า Precision โดยเฉลี่ยสูงเป็นลำดับที่ 2 คือวิธี L0L2Learn และเมื่อข้อมูลมีค่า SNR สูงเท่ากับ 20 แล้ววิธี L0L2Learn จะมีค่าเฉลี่ยของค่า Precision สูงกว่าวิธี L0Learn เล็กน้อย ส่วนวิธี L1 นั้น เมื่อ SNR มีค่าน้อย หรือข้อมูลมีสัญญาณรบกวนมากจะมีค่าเฉลี่ยของค่า Precision สูงกว่าวิธี A-L1 และ A-L1L2 ส่วนค่า Precision ที่ได้จากวิธี A-L1 และ A-L1L2 นั้นจะเห็นได้ว่ามีความใกล้เคียงกันเป็นอย่างมาก และมีค่าน้อยที่สุดเมื่อเปรียบเทียบกับวิธีอื่น ๆ ในกรณีที่ SNR มีค่าน้อย แต่เมื่อข้อมูลมีค่า SNR เพิ่มมากขึ้น จะพบว่าวิธี A-L1 และ A-L1L2 จะมีค่า SNR เพิ่มสูงขึ้น และเทียบเท่ากับวิธี L1 เมื่อ SNR มีค่าตั้งแต่ 10 ขึ้นไป

นอกจากนี้เมื่อเปรียบเทียบผลการวิเคราะห์ของข้อมูลที่ตัวแปรอิสระไม่มีความสัมพันธ์กัน กับกรณีที่ตัวแปรอิสระมีความสัมพันธ์กันปานกลางแล้ว โดยภาพรวมแล้วพบว่าวิธี L0L2Learn ยังคงมีค่า Precision สูงกว่าวิธีอื่น ยกเว้นกรณีที่ข้อมูลมีค่า SNR เท่ากับ 20 อีกทั้งยังพบว่าการวิเคราะห์ทั้ง 5 วิธี จะมีค่าเฉลี่ยของค่า Precision สูงกว่าเมื่อข้อมูลมีตัวแปรอิสระที่มีความสัมพันธ์กันในระดับปานกลาง เนื่องมากจากเมื่อข้อมูลมีความสัมพันธ์กันปานกลางแล้ว ตัวแปรอิสระที่ถูกคัดเลือกเข้าสู่ตัวแบบจะมีจำนวนลดลง ทำให้ค่าเฉลี่ยของค่า Precision เพิ่มสูงขึ้น

ตารางที่ 4.5 ค่าเฉลี่ยของค่า Precision จาก 100 ชุดข้อมูล เมื่อตัวแปรอิสระมีความสัมพันธ์กันปานกลาง

วิธี	ค่าสถิติ	ค่า SNR					
		0.1	0.5	1	5	10	20
n = 100, ρ = 0.5							
L0Learn	Mean	0.183	0.716	0.835	0.888	0.885	0.870
	SD.	0.337	0.329	0.216	0.141	0.153	0.145
L0L2Learn	Mean	0.146	0.491	0.533	0.725	0.828	0.916
	SD.	0.251	0.335	0.314	0.244	0.161	0.104
L1	Mean	0.112	0.407	0.360	0.371	0.331	0.335
	SD.	0.144	0.273	0.197	0.137	0.107	0.090
A-L1	Mean	0.051	0.107	0.141	0.289	0.349	0.411
	SD.	0.025	0.057	0.062	0.121	0.102	0.100
A-L1L2	Mean	0.051	0.103	0.142	0.301	0.351	0.412
	SD.	0.023	0.040	0.052	0.109	0.105	0.094



รูปที่ 4.5 Boxplot แสดงค่า Precision เมื่อตัวแปรอิสระมีความสัมพันธ์กันปานกลาง

3) เมื่อข้อมูลมีระดับ SNR แตกต่างกัน และตัวแปรอิสระมีความสัมพันธ์กันสูง (ρ = 0.9)

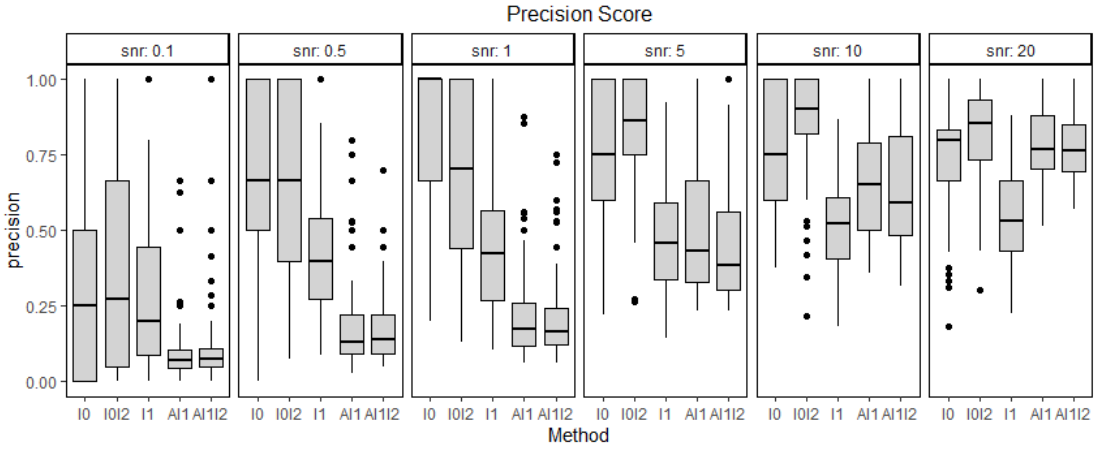
พิจารณาผลการวิเคราะห์จากตารางที่ 4.6 ซึ่งแสดงค่าเฉลี่ยของค่า Precision และรูปที่ 4.6 พบว่า เมื่อข้อมูลมีสัญญาณรบกวนสูง มีค่า SNR ต่ำวิธี L0Learn มีค่าเฉลี่ยของค่า Precision สูงเมื่อเปรียบเทียบกับวิธีอื่น ๆ แต่เมื่อ SNR มีค่ามากขึ้นและมีค่าตั้งแต่ 5 ขึ้นไปกลับพบว่าวิธี L0L2Learn จะมีค่าเฉลี่ยของค่า Precision สูงที่สุด ส่วนผลการวิเคราะห์ในวิธี L1 A-L1 และ A-L1L2 พบว่าเมื่อค่า SNR น้อยวิธี L1 จะมีค่าเฉลี่ยของค่า Precision มากกว่าวิธี A-L1 และ A-L1L2 แต่เมื่อ SNR มีค่าสูงขึ้นและมีค่าตั้งแต่ 5 ขึ้นไปวิธี A-L1 และ A-L1L2 จะมีค่าเฉลี่ยของค่า Precision มากกว่าวิธี L1 นอกจากนี้ยังพบว่าวิธี A-L1 และ A-L1L2 จะมีพฤติกรรมคล้ายคลึงกันเป็นอย่างมาก

นอกจากนี้เมื่อเปรียบเทียบผลการวิเคราะห์กับกรณีที่ข้อมูลมีตัวแปรอิสระที่ไม่มีความสัมพันธ์กัน และกรณีที่ตัวแปรอิสระมีความสัมพันธ์กันปานกลางแล้ว จะพบว่าค่าเฉลี่ยของค่า Precision ของกรณีนี้จะลดลงเล็กน้อยเมื่อเปรียบกับข้อมูลที่ ρ = 0.5 อย่างไรก็ตามพฤติกรรมการคัดเลือกตัวแปรของทั้ง 5 วิธียังคงมีลักษณะคล้ายเดิมคือวิธี L0Learn จะมีค่าเฉลี่ยของค่า Precision สูงเมื่อข้อมูลมีค่า SNR น้อย แต่เมื่อค่า SNR เพิ่มมากขึ้นและขณะเดียวกันตัวแปรอิสระมีความสัมพันธ์กันมากขึ้นวิธี L0L2Learn จะมีค่าเฉลี่ยของค่า Precision สูงที่สุด

ตารางที่ 4.6 ค่าเฉลี่ยของค่า Precision จาก 100 ชุดข้อมูล เมื่อตัวแปรอิสระมีความสัมพันธ์กันสูง

วิธี	ค่าสถิติ	ค่า SNR					
		0.1	0.5	1	5	10	20
n = 100, ρ = 0.9							
L0Learn	Mean	0.350	0.684	0.808	0.746	0.746	0.767
	SD.	0.388	0.272	0.235	0.209	0.192	0.183
L0L2Learn	Mean	0.392	0.682	0.700	0.841	0.875	0.827
	SD.	0.382	0.310	0.291	0.173	0.157	0.153
L1	Mean	0.306	0.439	0.446	0.474	0.527	0.538
	SD.	0.284	0.231	0.221	0.186	0.152	0.155
A-L1	Mean	0.095	0.183	0.210	0.494	0.662	0.785
	SD.	0.106	0.144	0.142	0.209	0.183	0.118
A-L1L2	Mean	0.111	0.170	0.210	0.452	0.647	0.779
	SD.	0.141	0.114	0.137	0.192	0.197	0.104

Chulalongkorn University



รูปที่ 4.6 Boxplot แสดงค่า Precision เมื่อตัวแปรอิสระมีความสัมพันธ์กันปานสูง

4.2.2 ค่าเฉลี่ยของค่าความระลึก (Average of Recall)

ค่าเฉลี่ยของค่าความระลึก (Average of Recall) เป็นการวัดความสามารถในการเรียกคืนตัวแปร โดยเฉลี่ยแล้วตัวแบบสามารถคัดเลือกตัวแปรได้ถูกต้องเป็นสัดส่วนเท่าใดเมื่อเปรียบเทียบกับจำนวนตัวแปรที่มีสัมประสิทธิ์ไม่เท่ากับศูนย์ที่แท้จริงทั้งหมด เหมาะกับการวิเคราะห์ที่ต้องการลดจำนวน False Negative หรือตัวแปรที่ตัวแบบทำนายให้มีค่าสัมประสิทธิ์เป็นศูนย์ แต่ในความเป็นจริงแล้วมีค่าสัมประสิทธิ์ไม่เป็นศูนย์ สำหรับผลการวิเคราะห์เมื่อข้อมูลมีระดับ SNR และตัวแปรอิสระไม่มีความสัมพันธ์กันในระดับต่าง ๆ มีรายละเอียด ดังนี้

1) เมื่อข้อมูลมีระดับ SNR แตกต่างกัน และไม่มีความสัมพันธ์กัน (ρ = 0.0)

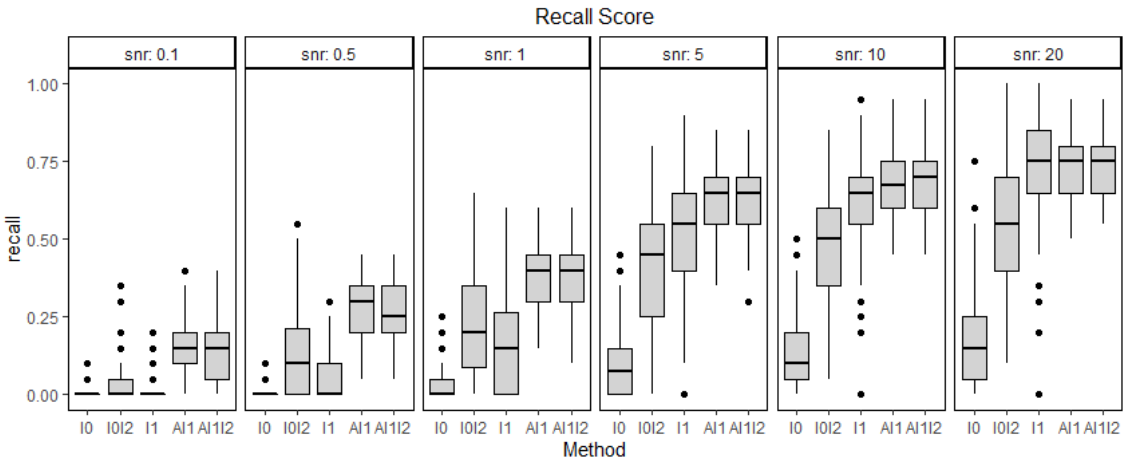
Chulalongkorn University

พิจารณาผลการวิเคราะห์จากตารางที่ 4.7 ซึ่งแสดงค่าเฉลี่ยของค่าความระลึก (Recall) และรูปที่ 4.7 จะเห็นได้ว่าวิธี A-L1 และ A-L1L2 เป็นวิธีที่มีค่าเฉลี่ยของค่าความระลึก (Recall) สูงใกล้เคียงกันมากและมีค่าสูงที่สุดในทุกกรณีแม้จะมีการเปลี่ยนแปลงของค่า SNR ด้านวิธี L1 นั้นเมื่อข้อมูลมีสัญญาณรบกวนสูง หรือมีค่า SNR ต่ำ วิธี L1 จะมีประสิทธิภาพในการเรียกคืนตัวแปรได้ไม่ดีนัก และมีประสิทธิภาพต่ำกว่าวิธี A-L1 , A-L1L2 และวิธี L0L2Learn แต่เมื่อค่า SNR เพิ่มมากขึ้นและมากกว่าเท่ากับ 5 จะพบว่าวิธี A-L1 จะมีค่าเฉลี่ยของค่า Recall มากกว่าวิธี L0L2Learn และหากค่า SNR เท่ากับ 20 วิธี L1 จะมีค่าเฉลี่ยของค่า Recall สูงใกล้เคียงกับวิธี A-L1 และ A-L1L2 ส่วนวิธี L0L2Learn เมื่อข้อมูลมีค่า SNR ต่ำจะมีประสิทธิภาพในการเรียกคืนตัวแปรสูงกว่าวิธี L1 เล็กน้อย แต่เมื่อค่า SNR มากขึ้นกลับพบว่าวิธี L0L2Learn มีประสิทธิภาพในการเรียกคืนตัวแปรน้อยกว่าวิธี L1 ส่วนวิธี L0Learn นั้นมีประประสิทธิภาพต่ำที่สุดในทุกกรณี ที่เป็นเช่นนี้เพราะวิธี L0Learn

เลือกตัวแปรเข้าสู่ตัวแบบน้อยมากเมื่อเปรียบเทียบกับวิธีอื่น ดังนั้นค่าเฉลี่ยของค่า Recall จึงมักจะต่ำกว่าวิธีอื่นเสมอ

ตารางที่ 4.7 ค่าเฉลี่ยของค่า Recall จาก 100 ชุดข้อมูล เมื่อตัวแปรอิสระไม่มีความสัมพันธ์กัน

วิธี	ค่าสถิติ	ค่า SNR					
		0.1	0.5	1	5	10	20
n = 100, ρ = 0.0							
L0Learn	Mean	0.004	0.005	0.023	0.098	0.124	0.172
	SD.	0.016	0.016	0.045	0.113	0.112	0.156
L0L2Learn	Mean	0.041	0.138	0.221	0.406	0.458	0.533
	SD.	0.072	0.135	0.161	0.189	0.174	0.211
L1	Mean	0.017	0.055	0.165	0.517	0.609	0.733
	SD.	0.038	0.079	0.152	0.208	0.187	0.169
A-L1	Mean	0.146	0.279	0.377	0.633	0.682	0.738
	SD.	0.086	0.088	0.104	0.101	0.090	0.104
A-L1L2	Mean	0.149	0.270	0.375	0.629	0.678	0.737
	SD.	0.089	0.092	0.102	0.105	0.088	0.099



รูปที่ 4.7 Boxplot แสดงค่า Recall เมื่อตัวแปรอิสระมีไม่ความสัมพันธ์กัน

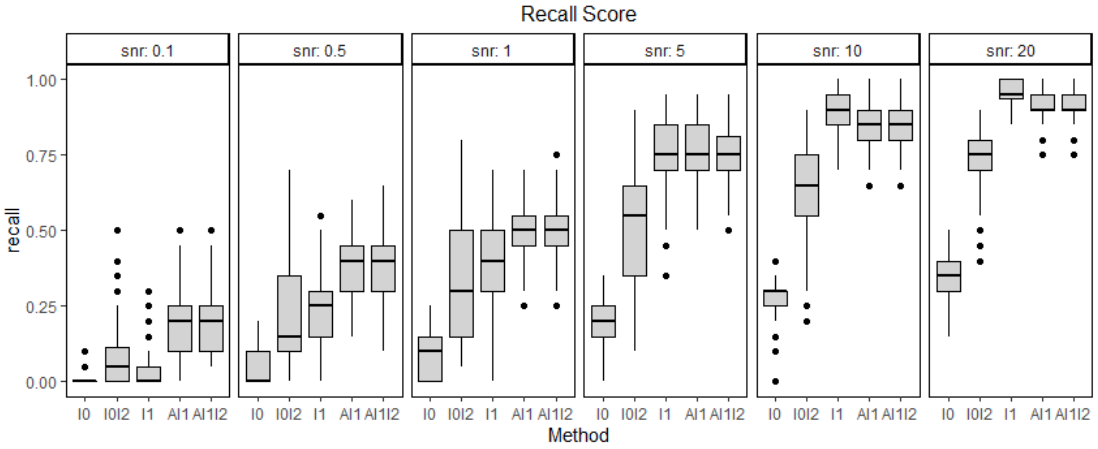
2) เมื่อข้อมูลมีระดับ SNR แตกต่างกัน และตัวแปรอิสระมีความสัมพันธ์กันปานกลาง (ρ = 0.5)

พิจารณาผลการวิเคราะห์จากตารางที่ 4.8 ซึ่งแสดงค่าเฉลี่ยของค่าความระลึก (Recall) และรูปที่ 4.8 จะเห็นได้ว่าเมื่อค่า SNR ต่ำหรือข้อมูลมีสัญญาณรบกวนสูง วิธี A-L1 และวิธี A-L1L2 จะให้ค่าเฉลี่ยของค่า Recall สูงกว่าวิธีอื่น ๆ อย่างเห็นได้ชัด และเมื่อค่า SNR มีค่าเพิ่มมากขึ้นหรือกล่าวได้ว่าข้อมูลมีสัญญาณรบกวนต่ำลง วิธี L1 จะมีค่าเฉลี่ยค่า Recall สูงขึ้นมากและหากค่า SNR มากกว่าหรือเท่ากับ10 วิธี L1 จะมีค่าเฉลี่ยของค่า Recall มากที่สุด หรือกว่าได้ว่าสามารถเรียกคืนตัวแปรที่ถูกต้องเข้าสู่ตัวแบบได้ดีที่สุด แต่ตัวแบบที่ได้จะมีจำนวนตัวแปรอิสระค่อนข้างมากประมาณ 60 ตัว ดังนั้นจึงส่งผลให้โอกาสที่จะดึงตัวแปรที่ถูกต้องเข้าสู่ตัวแบบได้ดีกว่าวิธีอื่น ๆ ที่มีจำนวนตัวแปรอิสระในตัวแบบน้อยกว่า ส่วนวิธี L0L2Learn จะมีประสิทธิภาพในการเรียกคืนตัวแปรไม่สูงนักเมื่อเปรียบเทียบกับวิธีอื่น ๆ แต่เมื่อค่า SNR มากขึ้นเท่ากับ 5 จะสังเกตได้ว่า วิเคราะห์ L0L2Learn จะมีค่าเฉลี่ยของค่า Recall เพิ่มสูงขึ้นอย่างเห็นได้ชัด แต่ยังคงมีประสิทธิภาพน้อยกว่าวิธี L1, A-L1 และ A-L1L2 ส่วนวิธี L0Learn นั้นมีประประสิทธิภาพต่ำที่สุดในทุกกรณี

นอกจากนี้เมื่อตัวแปรอิสระมีความสัมพันธ์กันปานกลาง (ρ = 0.5) จะสังเกตเห็นว่าวิธีวิเคราะห์ทั้ง 5 วิธีจะมีค่าเฉลี่ยของค่า Recall สูงกว่าข้อมูลที่ตัวแปรอิสระไม่มีความสัมพันธ์กัน นั่นคือเมื่อข้อมูลมีความสัมพันธ์กันปานกลางจะสามารถเรียกคืนตัวแปรได้ดีกว่าข้อมูลที่ไม่มีความสัมพันธ์กัน

ตารางที่ 4.8 ค่าเฉลี่ยของค่า Recall จาก 100 ชุดข้อมูล เมื่อตัวแปรอิสระมีความสัมพันธ์กันปานกลาง

วิธี	ค่าสถิติ	ค่า SNR					
		0.1	0.5	1	5	10	20
n = 100, ρ = 0.5							
L0Learn	Mean	0.006	0.039	0.083	0.220	0.275	0.338
	SD.	0.017	0.053	0.066	0.070	0.072	0.077
L0L2Learn	Mean	0.090	0.235	0.339	0.507	0.624	0.736
	SD.	0.111	0.180	0.196	0.189	0.141	0.100
L1	Mean	0.037	0.231	0.393	0.751	0.879	0.956
	SD.	0.063	0.138	0.134	0.110	0.072	0.046
A-L1	Mean	0.189	0.375	0.501	0.775	0.854	0.917
	SD.	0.099	0.100	0.103	0.095	0.078	0.053
A-L1L2	Mean	0.190	0.378	0.503	0.768	0.853	0.916
	SD.	0.098	0.108	0.105	0.092	0.078	0.055



รูปที่ 4.8 Boxplot แสดงค่า Recall เมื่อตัวแปรอิสระมีความสัมพันธ์กันปานกลาง

3) เมื่อข้อมูลมีระดับ SNR แตกต่างกัน และตัวแปรอิสระมีความสัมพันธ์กันสูง (ρ = 0.9)

พิจารณาผลการวิเคราะห์จากตารางที่ 4.9 ซึ่งแสดงค่าเฉลี่ยของค่าความระลึก (Recall) และรูปที่ 4.9 พบว่าวิธี A-L1L2 มีค่าเฉลี่ยของค่า Recall สูงที่สุดในแม้ว่าข้อมูลจะมีระดับ SNR น้อยหรือระดับ SNR สูง และเป็นวิธีที่มีประสิทธิภาพในการเรียกคืนตัวแปรได้ดีที่สุดเมื่อเปรียบเทียบกับวิธีอื่น ๆ ที่มีประสิทธิภาพรองลงมาเป็นลำดับที่ 2 และ 3 คือวิธี A-L1 และ L1 ตามลำดับ แต่เมื่อค่า SNR สูงขึ้นหรือข้อมูลมีสัญญาณรบกวนน้อย วิธี A-L1 และ L1 จะมีประสิทธิภาพเพิ่มมากขึ้นอย่างมาก และมีประสิทธิภาพเทียบเท่ากับวิธี A-L1L2 เมื่อข้อมูลมีระดับ SNR เท่ากับ 20 จะเห็นได้ว่าเมื่อค่า SNR ต่ำหรือข้อมูลมีสัญญาณรบกวนสูง วิธี A-L1 และวิธี A-L1L2 จะให้ค่าเฉลี่ยของค่า Recall สูงกว่าวิธีอื่น ๆ อย่างเห็นได้ชัด และเมื่อค่า SNR มีค่าเพิ่มมากขึ้นหรือข้อมูลมีสัญญาณรบกวนต่ำลง วิธี L1 จะมีค่าเฉลี่ยค่า Recall สูงขึ้นมากและหากค่า SNR มากกว่าหรือเท่ากับ 20 วิธี L1 จะมีค่าเฉลี่ยของค่า Recall มากที่สุด หรือกว่าได้ว่าสามารถเรียกคืนตัวแปรที่ถูกต้องเข้าสู่ตัวแบบได้ดีที่สุด อย่างไรก็ตามตัวแบบที่ได้จะมีจำนวนตัวแปรอิสระค่อนข้างมากประมาณ 60 ตัว ดังนั้นจึงส่งผลให้โอกาสที่จะดึงตัวแปรที่ถูกต้องเข้าสู่ตัวแบบได้ดีกว่าวิธีอื่น ๆ ที่มีจำนวนตัวแปรอิสระในตัวแบบน้อยกว่าส่วนวิธี L0L2Learn และวิธี L0Learn จะมีประสิทธิภาพในการเรียกคืนตัวแปรเป็นลำดับที่ 4 และ 5 ตามลำดับ

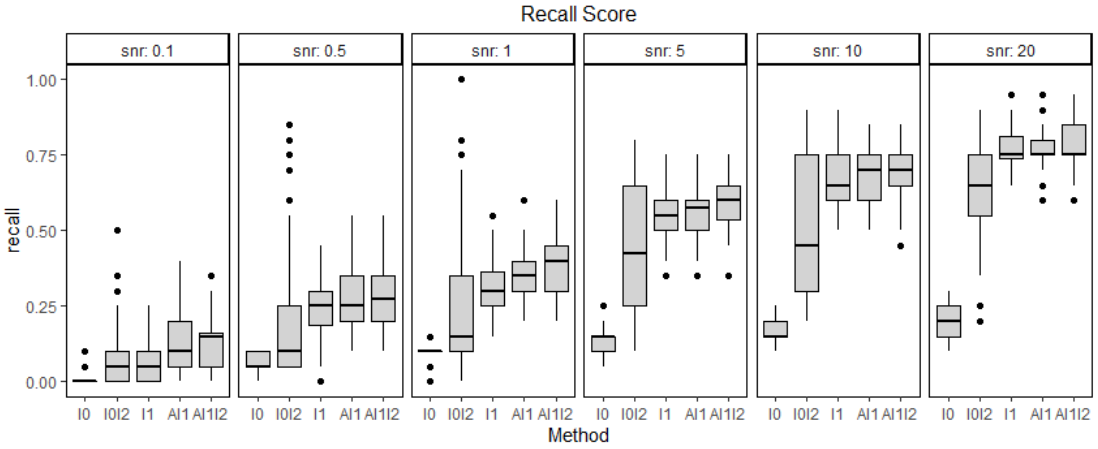
จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

นอกจากนี้เมื่อเปรียบเทียบกับกรณีที่ข้อมูลมีตัวแปรอิสระไม่มีความสัมพันธ์กันและข้อมูลที่ตัวแปรอิสระมีความสัมพันธ์กันปานกลาง พบว่าทุกวิธีจะมีความสามารถในการเรียกคืนตัวแปรต่ำลงเมื่อเปรียบเทียบกับค่าเฉลี่ยของค่า Recall ที่ได้จากกรณีที่ตัวแปรอิสระมีความสัมพันธ์ปานกลาง และมีค่าใกล้เคียงกับกรณีที่ตัวแปรอิสระไม่มีความสัมพันธ์กัน อย่างไรก็ตามเมื่อเปรียบเทียบพฤติกรรมการคัดเลือกตัวแปรของทั้ง 5 วิธี พบว่ากรณีที่ตัวแปรอิสระไม่มีความสัมพันธ์กัน (ρ = 0.0) และกรณีที่ตัวแปรอิสระมีความสัมพันธ์กันสูง (ρ = 0.9) ทั้งสองวิธีมีผลลัพธ์ใกล้เคียงกันเป็นอย่างมาก คือ หาก SNR มีค่าน้อยวิธี A-L1 และวิธี A-L1L2 จะให้ค่า Recall สูงกว่าวิธีอื่น ๆ แต่เมื่อค่า SNR เพิ่มมากขึ้นเท่ากับ 20 วิธี L1 จะมีค่าเฉลี่ยของค่า Recall เทียบเท่ากับกับวิธี A-L1 และวิธี A-L1L2 แต่ในกรณีที่ข้อมูลมีความสัมพันธ์กันปานกลางและมีค่า SNR มากกว่าหรือเท่ากับ 5 วิธี L1 จะมีค่าเฉลี่ยของค่า Recall มากที่สุด

ตารางที่ 4.9 ค่าเฉลี่ยของค่า Recall จาก 100 ชุดข้อมูล เมื่อตัวแปรอิสระมีความสัมพันธ์กันสูง

วิธี	ค่าสถิติ	ค่า SNR					
		0.1	0.5	1	5	10	20
n = 100, ρ = 0.9							
L0Learn	Mean	0.012	0.059	0.089	0.141	0.163	0.195
	SD.	0.024	0.039	0.031	0.035	0.037	0.044
L0L2Learn	Mean	0.075	0.195	0.258	0.438	0.502	0.624
	SD.	0.110	0.208	0.218	0.225	0.216	0.161
L1	Mean	0.066	0.234	0.313	0.560	0.664	0.773
	SD.	0.067	0.088	0.093	0.076	0.082	0.071
A-L1	Mean	0.126	0.280	0.354	0.568	0.676	0.771
	SD.	0.088	0.092	0.086	0.079	0.077	0.068
A-L1L2	Mean	0.134	0.290	0.376	0.577	0.682	0.781
	SD.	0.084	0.092	0.094	0.076	0.083	0.075

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University



รูปที่ 4.9 Boxplot แสดงค่า Recall เมื่อตัวแปรอิสระมีความสัมพันธ์กันสูง

4.2.3 ค่าเฉลี่ยของค่าพื้นที่ใต้กราฟ ROC (Average of AUC Score)

จากตัววัดประสิทธิภาพในการคัดเลือกตัวแบบด้วยค่าเฉลี่ยของค่า Recall และค่าเฉลี่ยของค่า Precision ที่ผ่านมาจะเห็นได้ว่าตัวแบบที่มีค่าเฉลี่ยของค่า Recall สูงมักจะมีค่าเฉลี่ยของค่า Precision ต่ำ และในทางกลับกับตัวแบบที่ทำให้ผลการคัดเลือกมีค่าเฉลี่ยของค่า Precision สูงก็จะมีค่าเฉลี่ยของค่า Recall ต่ำ โดยทั่วไปจะต้องเลือกอย่างใดอย่างหนึ่งเท่านั้น ในการวิจัยจึงได้ใช้ค่าเฉลี่ยของค่าพื้นที่ใต้กราฟ ROC (Average of AUC Score) เป็นตัววัดประสิทธิภาพตัวหนึ่ง ซึ่งจะช่วยบ่งบอกได้ว่าตัวแบบมีความสามารถในการแยกตัวแปรอิสระที่มีค่าสัมประสิทธิ์ไม่เท่ากับศูนย์ออกจากตัวแปรอิสระอื่น ๆ ที่มีค่าสัมประสิทธิ์เป็นศูนย์ได้ดีเพียงใด ซึ่งผลการวิเคราะห์ในแต่ละกรณีมีรายละเอียดดังนี้

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

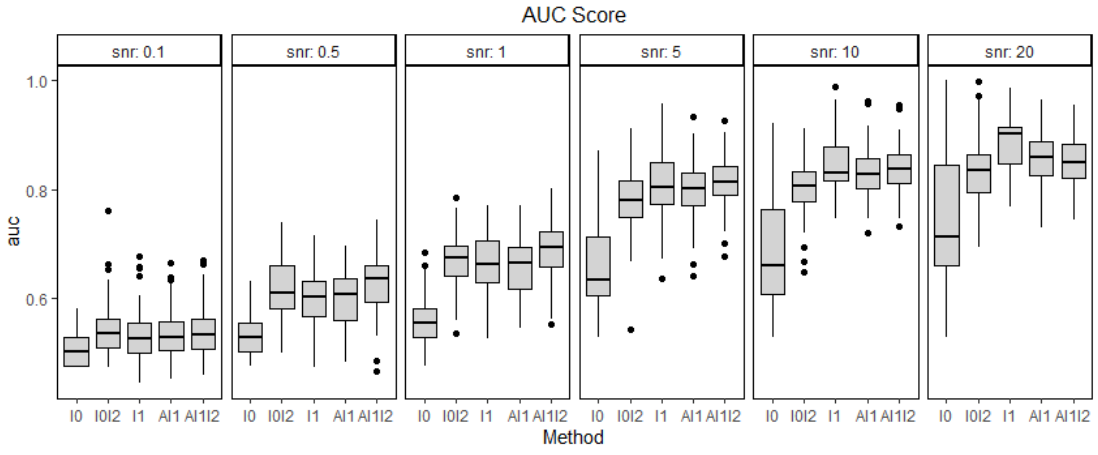
1) เมื่อข้อมูลมีระดับ SNR แตกต่างกัน และตัวแปรอิสระไม่มีความสัมพันธ์กัน (ρ = 0.0)

พิจารณาผลการวิเคราะห์จากรูปที่ 4.10 และตารางแสดงค่าสถิติของผลการวิเคราะห์ในตารางที่ 4.10 จะเห็นได้ว่าวิธี A-L1L2 มีค่าเฉลี่ยของค่า AUC สูงที่สุดในทุกระดับ SNR ยกเว้นกรณีที่ SNR มีค่าเท่ากับ 20 หมายความว่าแม้ข้อมูลจะมีสัญญาณรบกวนมากหรือน้อยวิธี A-L1L2 ยังมีประสิทธิภาพในการคัดเลือกตัวแปรเข้าสู่ตัวแบบเสมอ ส่วนวิธี L0L2Learn ในกรณีที่ข้อมูลมีค่า SNR ต่ำหรือข้อมูลมีสัญญาณรบกวนสูงแล้ว วิธี L0L2Learn จะมีค่าเฉลี่ยของค่า AUC สูงรองจากวิธี A-L1L2 แต่หากข้อมูลมีระดับ SNR เพิ่มมากขึ้นการวิเคราะห์จะมีความสามารถในการคัดแยกตัวแปรได้ดีขึ้นหรือมีค่าเฉลี่ยของค่า AUC สูงขึ้นในทุก ๆ วิธี แต่เมื่อ SNR มีค่าตั้งแต่ 5 แล้ว วิธี L0L2Learn จะมีค่าเฉลี่ยของค่า AUC น้อยกว่าวิธี A-L1 และ A-L1L2 หมายความว่าเมื่อข้อมูลมีสัญญาณรบกวน

น้อยแล้ววิธี A-L1 และ A-L1L2 จะสามารถคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบได้ดีว่าวิธี L0L2Learn และหาก SNR เพิ่มสูงขึ้นและมีค่ามากกว่าเท่ากับ 10 แล้ววิธี A-L1 และ A-L1L2 จะมีประสิทธิภาพเทียบเท่ากับวิธี A-L1L2 ส่วนวิธี L0Learn นั้นมีค่าเฉลี่ยของค่า AUC น้อยที่สุดทุกกรณี

ตารางที่ 4.10 ค่าเฉลี่ยของค่า AUC Score จาก 100 ชุดข้อมูล เมื่อตัวแปรอิสระไม่มีความสัมพันธ์กัน

วิธี	ค่าสถิติ	ค่า SNR					
		0.1	0.5	1	5	10	20
n = 100, ρ = 0.0							
L0Learn	Mean	0.509	0.530	0.558	0.654	0.683	0.749
	SD.	0.029	0.035	0.043	0.077	0.098	0.129
L0L2Learn	Mean	0.543	0.616	0.672	0.782	0.802	0.833
	SD.	0.045	0.051	0.046	0.057	0.047	0.064
L1	Mean	0.530	0.599	0.661	0.806	0.845	0.885
	SD.	0.047	0.050	0.054	0.056	0.047	0.054
A-L1	Mean	0.531	0.602	0.659	0.799	0.827	0.856
	SD.	0.044	0.047	0.053	0.053	0.046	0.051
A-L1L2	Mean	0.540	0.627	0.690	0.815	0.836	0.852
	SD.	0.047	0.050	0.055	0.044	0.042	0.044



รูปที่ 4.10 Boxplot แสดงค่า AUC Score เมื่อตัวแปรอิสระไม่มีความสัมพันธ์กัน

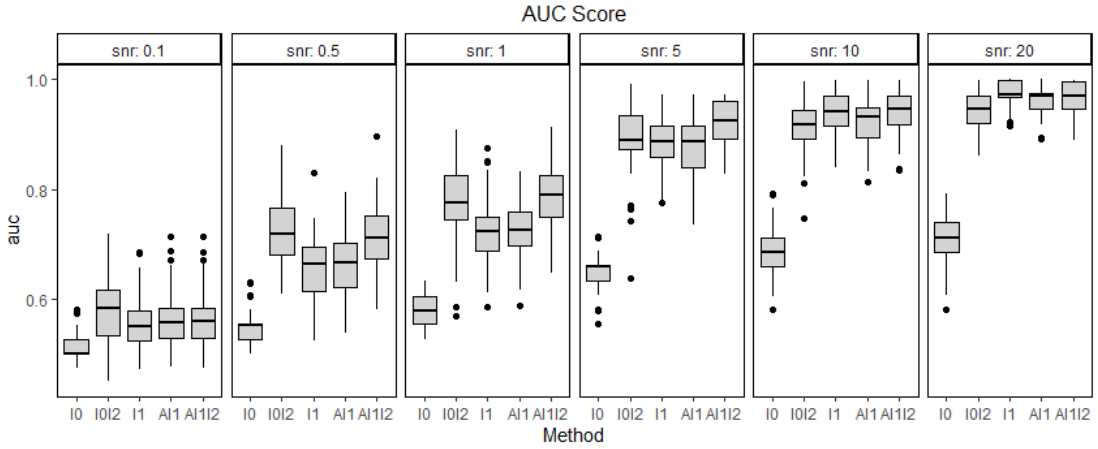
2) เมื่อข้อมูลมีระดับ SNR แตกต่างกัน และตัวแปรอิสระมีความสัมพันธ์กันปานกลาง (ρ = 0.5)

พิจารณาผลการวิเคราะห์จากรูปที่ 4.11 และตารางแสดงค่าสถิติของผลการวิเคราะห์ในตารางที่ 4.11 พบว่าหากข้อมูลมีค่า SNR น้อยกว่าหรือเท่ากับ 1 วิธี L0L2Learn จะมีค่าเฉลี่ยของค่า AUC สูงที่สุด หมายความว่าเมื่อข้อมูลมีสัญญาณรบกวนมากวิธี L0L2Learn สามารถคัดแยกตัวแปรมีค่าสัมประสิทธิ์ไม่เท่ากับศูนย์ออกจากตัวแปรอิสระได้ดีกว่าวิธีอื่น ๆ และเมื่อข้อมูลมีค่า SNR มากขึ้นทุก ๆ วิธีจะมีค่าเฉลี่ยของ AUC เพิ่มสูงขึ้น แต่เมื่อค่า SNR เพิ่มสูงถึง 1 แล้ววิธี A-L1L2 จะมีค่าเฉลี่ยของค่า AUC สูงเทียบเท่ากับวิธี L0L2Learn และหากค่า SNR มีค่าตั้งแต่ 5 ขึ้นไปวิธี A-L1L2 จะสามารถคัดแยกตัวแปรอิสระได้ดีกว่าวิธี L0L2Learn และเมื่อข้อมูลมีค่า SNR สูงเท่ากับ 20 วิธี L1 สามารถคัดเลือกตัวแปรได้ดี ส่วนในวิธี L0Learn นั้นมีค่าเฉลี่ยของค่า AUC น้อยที่สุดทุกกรณี

นอกจากนี้เมื่อเปรียบเทียบการวิเคราะห์กับกรณีที่ข้อมูลมีตัวแปรอิสระไม่มีความสัมพันธ์กันจะเห็นได้ว่าข้อมูลที่มีความสัมพันธ์กันปานกลางจะมีค่าเฉลี่ยของค่า AUC สูงกว่าในทุก ๆ วิธีการวิเคราะห์และทุกกรณีแม้จะมีการปรับเปลี่ยนค่า SNR หมายความว่าหากตัวแปรอิสระมีความสัมพันธ์กันปานกลางทั้ง 5 วิธีจะสามารถคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบได้ดีขึ้น โดยเฉพาะในกรณีที่ข้อมูลมี SNR น้อยกว่าหรือเท่ากับ 1 และข้อมูลมีความสัมพันธ์กันปานกลางวิธี L0L2Learn จะมีประสิทธิภาพในการคัดเลือกตัวแบบได้ดี

ตารางที่ 4.11 ค่าเฉลี่ยของค่า AUC Score จาก 100 ชุดข้อมูล เมื่อตัวแปรอิสระมีความสัมพันธ์กันปานกลาง

วิธี	ค่าสถิติ	ค่า SNR					
		0.1	0.5	1	5	10	20
n = 100, ρ = 0.5							
L0Learn	Mean	0.513	0.548	0.575	0.649	0.679	0.706
	SD.	0.025	0.028	0.029	0.033	0.041	0.047
L0L2Learn	Mean	0.582	0.720	0.777	0.892	0.922	0.949
	SD.	0.058	0.058	0.071	0.052	0.042	0.032
L1	Mean	0.558	0.658	0.724	0.884	0.937	0.974
	SD.	0.043	0.055	0.054	0.044	0.035	0.023
A-L1	Mean	0.562	0.663	0.729	0.880	0.929	0.960
	SD.	0.046	0.054	0.050	0.048	0.039	0.028
A-L1L2	Mean	0.562	0.711	0.786	0.922	0.948	0.970
	SD.	0.046	0.059	0.057	0.039	0.034	0.026



รูปที่ 4.11 Boxplot แสดงค่า AUC Score เมื่อตัวแปรอิสระมีความสัมพันธ์กันปานกลาง

3) เมื่อข้อมูลมีระดับ SNR แตกต่างกัน และตัวแปรอิสระมีความสัมพันธ์กันสูง (ρ = 0.9)

พิจารณาผลการวิเคราะห์จากรูปที่ 4.12 และตารางแสดงค่าสถิติของผลการวิเคราะห์ในตารางที่ 4.12 พบว่าวิธี L0L2Learn มีค่าเฉลี่ยของค่า AUC สูงที่สุดในทุกระดับ SNR หมายความว่าแม้ข้อมูลจะมีสัญญาณรบกวนมากหรือน้อยวิธี L0L2Learn ยังมีประสิทธิภาพในการคัดเลือกตัวแปรเข้าสู่ตัวแบบเสมอเมื่อข้อมูลมีตัวแปรอิสระที่มีความสัมพันธ์กันสูง ส่วนวิธี A-L1 และ A-L1L2 พบว่าในชุดข้อมูลที่มีค่า SNR ต่ำวิธี A-L1 และ A-L1L2 จะมีค่าเฉลี่ยของค่า AUC ใกล้เคียงกันมาก แต่เมื่อข้อมูลมีค่า SNR เพิ่มสูงขึ้น และมีค่ามากกว่าเท่ากับ 1 แล้ววิธี A-L1L2 จะมีค่าเฉลี่ยของค่า AUC สูงกว่าวิธี A-L1 และยิ่ง SNR มีค่ามากขึ้นเท่าใดวิธี A-L1L2 ยิ่งมีประสิทธิภาพในการคัดเลือกตัวแปรเข้าสู่ตัวแบบสูงขึ้นเท่านั้น เห็นได้ว่าเมื่อค่า SNR มากกว่าเท่ากับ 10 วิธี A-L1L2 จะมีค่าเฉลี่ยของค่า AUC สูงเทียบเท่ากับวิธี L0L2Learn ส่วนวิธี L1 มีค่าเฉลี่ยของค่า AUC น้อยเป็นลำดับที่ 4 และวิธี L0Learn นั้นมีค่าเฉลี่ยของค่า AUC น้อยเป็นลำดับที่ 5 ในทุก ๆ ค่า SNR

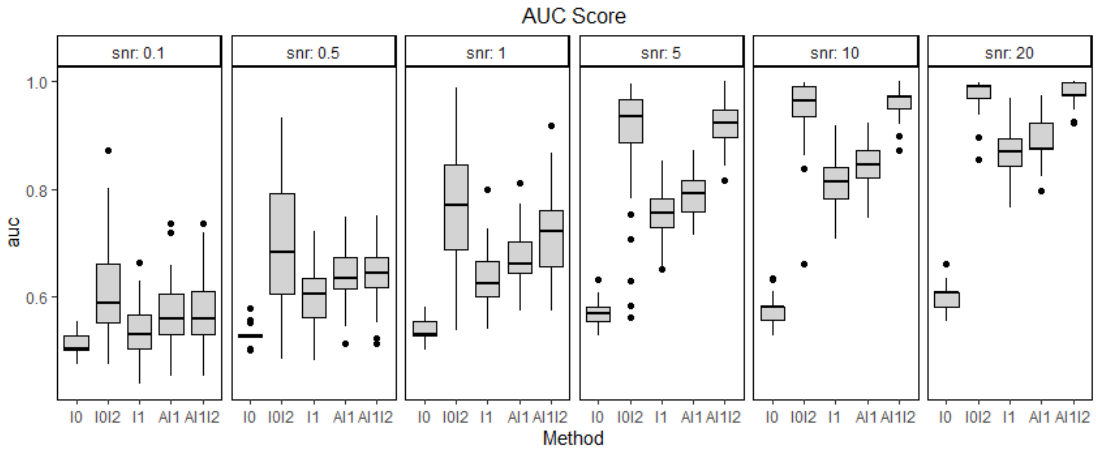
นอกจากนี้เมื่อเปรียบเทียบการวิเคราะห์กับกรณีที่ข้อมูลมีตัวแปรอิสระมีความสัมพันธ์กันปานกลาง พบว่าข้อมูลที่มีความสัมพันธ์กันสูงจะมีค่าเฉลี่ยของค่า AUC น้อยกว่าข้อมูลที่มีความสัมพันธ์กันในระดับปานกลาง ยกเว้นวิธี L0L2Learn และวิธี A-L1L2 จะมีค่าเฉลี่ยของค่า AUC สูงแต่ข้อมูลต้องมีค่า SNR สูงด้วย และเมื่อเปรียบเทียบพฤติกรรมการคัดเลือกตัวแปรเข้าสู่ตัวแบบของแต่ละวิธีแล้วจะสังเกตได้ว่าหากข้อมูลมีตัวแปรอิสระไม่สัมพันธ์กันวิธี A-L1L2 จะทำงานได้เป็นอย่างดี หากตัวแปรอิสระเริ่มมีความสัมพันธ์กันแล้ววิธี A-L1L2 จะมีประสิทธิภาพก็ต่อเมื่อข้อมูลมีค่า SNR สูง ส่วนวิธี L0L2Learn นั้นกลับพบว่าหากตัวแปรอิสระไม่มีความสัมพันธ์กันแล้วจะคัดเลือกตัวแปรเข้าสู่ตัวแบบได้ไม่ดีนัก แต่เมื่อตัวแปรอิสระเริ่มมีความสัมพันธ์กันมากขึ้นวิธี L0L2Learn ดีที่สุดก็

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

ต่อเมื่อข้อมูลมีค่า SNR น้อย แต่หากตัวแปรอิสระเริ่มมีความสัมพันธ์กันเพิ่มสูงขึ้น จนเท่ากัน 0.9 แล้ว วิธี L0L2Learn จะมีประสิทธิภาพในการคัดเลือกตัวแปรเข้าสู้ตัวแบบมากที่สุดแม้ว่าค่า SNR จะเปลี่ยนแปลงไป

ตารางที่ 4.12 ค่าเฉลี่ยของค่า AUC Score จาก 100 ชุดข้อมูล เมื่อตัวแปรอิสระมีความสัมพันธ์กันสูง

วิธี	ค่าสถิติ	ค่า SNR					
		0.1	0.5	1	5	10	20
n = 100, ρ = 0.9							
L0Learn	Mean	0.511	0.528	0.538	0.568	0.578	0.597
	SD.	0.021	0.018	0.019	0.021	0.020	0.021
L0L2Learn	Mean	0.608	0.699	0.765	0.912	0.955	0.982
	SD.	0.077	0.113	0.106	0.087	0.049	0.021
L1	Mean	0.536	0.599	0.633	0.754	0.812	0.874
	SD.	0.042	0.046	0.047	0.043	0.046	0.038
A-L1	Mean	0.566	0.640	0.674	0.783	0.842	0.891
	SD.	0.052	0.046	0.047	0.035	0.038	0.037
A-L1L2	Mean	0.568	0.642	0.718	0.919	0.962	0.982
	SD.	0.053	0.048	0.072	0.037	0.028	0.021



รูปที่ 4.12 Boxplot แสดงค่า AUC Score เมื่อตัวแปรอิสระมีความสัมพันธ์กันสูง

บทที่ 5
สรุปและอภิปรายผลการวิจัย

การวิจัยครั้งนี้เป็นการวิจัยเชิงทดลอง มีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าสัมประสิทธิ์การถดถอยเชิงเส้นสำหรับข้อมูลที่มีมิติสูง ด้วยวิธี Penalized regression ทั้งหมด 5 วิธี คือ 1) L0Learn 2) L0L2Learn 3) L1 4) A-L1 และ 5) A-L1L2 ในการเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าสัมประสิทธิ์การถดถอยเชิงเส้น จะทำการเปรียบเทียบผลลัพธ์โดยมุ่งเน้นไปที่ 2 กรณีคือ 1) ตัวแปรอิสระมีสหสัมพันธ์กันเท่ากับ 0.0, 0.5 และ 0.9 และ 2) เมื่อข้อมูลมีค่า SNR = 0.1, 0.5, 1, 5, 10 และ 20 นั่นคือในแต่ละสถานการณ์ข้อมูลจะมีสัญญาณรบกวนแตกต่างกัน โดยแต่ละกรณีจะทำการจำลองข้อมูลที่มีขนาดตัวอย่างเท่ากับ 100 ตัวอย่าง โดยการเปรียบเทียบประสิทธิภาพจะเปรียบเทียบใน 2 ด้าน คือ 1) เปรียบเทียบประสิทธิภาพด้านการพยากรณ์ ซึ่งพิจารณาจากความถูกต้องของการประมาณค่าสัมประสิทธิ์การถดถอยซึ่งวัดจากค่าเฉลี่ยของค่าคลาดเคลื่อนการทำนาย (AMSE) และ 2) ความถูกต้องในการคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบ ซึ่งพิจารณาจากค่าเฉลี่ยของค่า Precision Recall และค่า AUC ซึ่งได้จากการจำลองข้อมูลซ้ำจำนวน 100 รอบในแต่ละสถานการณ์

5.1 สรุปผลการวิจัย

ผลการเปรียบเทียบการวิเคราะห์แบ่งเป็น 2 ด้าน ด้านที่ 1 คือประสิทธิภาพในการพยากรณ์ของตัวแบบซึ่งแสดงผลการวิเคราะห์ค่าเฉลี่ยของค่าความคลาดเคลื่อนกำลังสอง (ซึ่งอยู่ในรูป Logarithms เนื่องจากต้องการลดขนาดความคลลาดเคลื่อนลงเพื่อให้สามารถเห็นความแตกต่างของแต่ละวิธีได้อย่างชัดเจนมากขึ้น) ในรูปที่ 5.1 พบว่า เมื่อชุดข้อมูลมีตัวแปรอิสระไม่มีความสัมพันธ์กัน วิธี L1 จะมีประสิทธิภาพในการพยากรณ์ในการพยากรณ์สูงที่สุดในทุก ๆ ค่า SNR ส่วนวิธี L0L2Learn และวิธี L0Learn จะมีประสิทธิภาพในการพยากรณ์รองลงมาจากวิธี L1 เมื่อข้อมูลมีค่า SNR ต่ำและก็มีค่า AMSE น้อยกว่าวิธี L1 ไม่มากนัก แต่เมื่อข้อมูลมีค่า SNR เพิ่มสูงขึ้นวิธี L0Learn จะมีประสิทธิภาพในการพยากรณ์น้อยกว่าวิธีอื่น ๆ อย่างเห็นได้ชัด ส่วนวิธี A-L1 และวิธี A-L1L2 จะมีความสามารถในการพยากรณ์เพิ่มสูงขึ้น และมีประสิทธิภาพเทียบเคียงกับวิธี L1 แต่ในกรณีที่ข้อมูลมีตัวแปรอิสระที่มีความสัมพันธ์กันในระดับปานกลางทั้ง 5 วิธีจะมีพฤติกรรมคล้ายคลึงกับผลการวิเคราะห์ในกรณีที่ตัวแปรอิสระมีความสัมพันธ์กันสูง คือ หากข้อมูลมีค่า SNR ต่ำวิธี L1 วิธี L0L2Learn และวิธี L0Learn จะมีประสิทธิภาพในการพยากรณ์สูงเป็นลำดับที่ 1 2 และ 3 ตามลำดับ แต่เมื่อข้อมูลมีค่า SNR เพิ่มสูงขึ้นและมีค่าตั้งแต่ 1 ขึ้นไปวิธี L0Learn จะมีประสิทธิภาพ

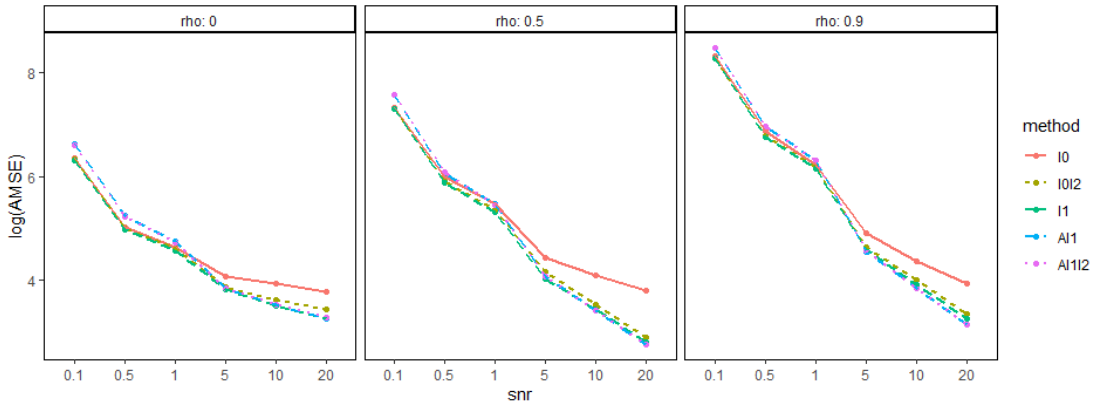
ในการพยากรณ์ต่ำที่สุดเมื่อเปรียบเทียบกับวิธีอื่น ๆ ส่วนวิธี A-L1 และวิธี A-L1L2 จะมีประสิทธิภาพในการพยากรณ์เพิ่มสูงขึ้น และสูงที่สุดเมื่อข้อมูลมีค่า SNR ตั้งแต่ 10 และ 5 ขึ้นไปเมื่อตัวแปรอิสระมีความสัมพันธ์กันปานกลางและตัวแปรอิสระมีความสัมพันธ์กันสูงตามลำดับ

ผลการเปรียบด้านความสามารถในการคัดเลือกตัวแปรเข้าสู่ตัวแบบโดยพิจารณาจากค่าเฉลี่ยของค่า Precision Recall และ AUC Score เมื่อพิจารณาด้วยค่าเฉลี่ยของค่า Precision จากรูปที่ 5.2 จะเห็นได้ว่าเมื่อค่า SNR น้อยทุกวิธีการจะมีค่าเฉลี่ยของค่า Precision ค่อนข้างต่ำ แต่วิธี L0Learn วิธี L1 และวิธี L0L2Learn จะมีประสิทธิภาพมากกว่าวิธี วิธี A-L1 วิธี A-L1L2 แต่เมื่อ SNR มีค่ามากขึ้นวิธี L0Learn และวิธี L0L2Learn จะมีประสิทธิภาพดีกว่าวิธีอื่น ๆ อย่างเห็นได้ชัด และเมื่อพิจารณาจากค่าเฉลี่ยของค่า Recall จากรูปที่ 5.3 พบว่าหากข้อมูลมีค่า SNR ต่ำและตัวแปรอิสระไม่มีความสัมพันธ์กันวิธี A-L1 และวิธี A-L1L2 จะมีค่าเฉลี่ยของค่า Recall สูงที่สุด รองลงมาคือ วิธี L0L2Learn วิธี L1 และวิธี L0Learn ตามลำดับ แต่หากข้อมูลมีค่า SNR เพิ่มมากขึ้นวิธี L1 จะมีประสิทธิภาพมากกว่าวิธี L0L2Learn และมีประสิทธิภาพมากที่สุดเมื่อข้อมูลมีค่า SNR มากกว่าหรือเท่ากับ 10 และเมื่อพิจารณาจากค่าเฉลี่ยของค่า AUC Score ซึ่งสามารถบ่งชี้ความสามารถทั้งด้านความแม่นยำในการคัดเลือกตัวแปรและในขณะเดียวกันก็บอกถึงความสามารถในการเรียกคืนตัวแปรได้ จากผลลัพธ์ซึ่งแสดงในรูปที่ 5.4 พบว่า หากข้อมูลที่ตัวแปรอิสระไม่มีความสัมพันธ์กันและกรณีที่ตัวแปรอิสระมีความสัมพันธ์กันปานกลางวิธี L1 วิธี A-L1 วิธี A-L1L2 และวิธี L0L2Learn มีความสามารถเลือกตัวแปรเข้าสู่ตัวแบบสูงใกล้เคียงกันแม้จะมีการเปลี่ยนแปลงค่า SNR แต่เมื่อข้อมูลมีตัวแปรอิสระที่มีความสัมพันธ์กันสูงเท่ากับ 0.9 (ρ = 0.9) และข้อมูลมีระดับ SNR ต่ำวิธี L0L2Learn จะเป็นวิธีที่มีประสิทธิภาพด้านการคัดเลือกตัวแปรเข้าสู่ตัวแบบที่โดดเด่นที่สุด และเมื่อข้อมูลมีค่า SNR เพิ่มสูงขึ้นวิธี A-L1L2 และวิธี L0L2Learn จะเป็นวิธีที่มีประสิทธิภาพโดดเด่นเท่าเทียมกัน ส่วนวิธี L0Learn เป็นวิธีที่มีประสิทธิภาพในการคัดเลือกตัวแปรเข้าสู่ตัวแบบน้อยกว่าวิธีอื่น ๆ ในทุกกรณี

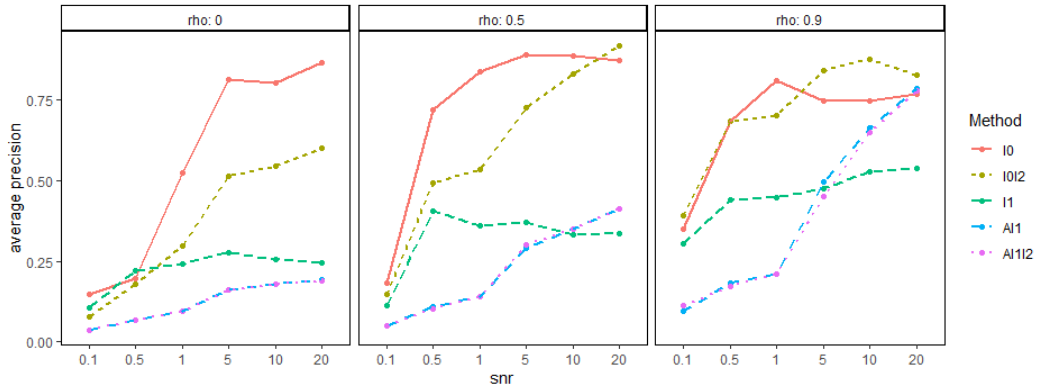
นอกจากการวัดประสิทธิภาพในการพยากรณ์และประสิทธิภาพในการคัดเลือกตัวแปลเข้าสู่ตัวแบบซึ่งเป็นตัววัดประสิทธิภาพที่สำคัญแล้ว ผู้วิจัยยังได้ศึกษาเพิ่มเติมถึงขนาดของตัวแบบที่ได้จากการวิเคราะห์ทั้ง 5 วิธีเพิ่มเติมด้วย เนื่องจากจำนวนตัวแปลที่อยู่ในตัวแบบ (Support Size) มีผลต่อต้นทุนในการวิเคราะห์ (Computation Cost) เวลาที่ใช้ในการวิเคราะห์ รวมไปถึงความสามารถในการอธิบายตัวแบบ (Interpretability) ของตัวแบบด้วย จากผลการวิเคราะห์ซึ่งแสดงในรูปที่ 5.3 พบว่าวิธี L1 ซึ่งมีประสิทธิภาพด้านการพยากรณ์โดดเด่นโดยเฉพาะในกรณีที่ชุดข้อมูลมีตัวแปรอิสระที่ไม่สัมพันธ์กันและมีค่า SNR ต่ำนั้น จะมีการคัดเลือกตัวแปรเข้าสู่ตัวแบบน้อยเมื่อข้อมูลมีค่า SNR ต่ำ แต่เมื่อข้อมูลมีค่า SNR มากขึ้นจะมีจำนวนตัวแปรที่อยู่ในตัวแบบมากขึ้นส่วนในตัวแบบที่ได้จากวิธี A-L1L2 และวิธี A-L1 ซึ่งเป็นวิธีที่มีประสิทธิภาพโดดเด่นในการคัดเลือกตัวแปรเข้าสู่ตัวแบบพบว่าใน

ชุดข้อมูลที่มีค่า SNR ต่ำจะมีจำนวนตัวแปรที่อยู่ในตัวแบบค่อนข้างมาก แต่เมื่อค่า SNR สูงขึ้นตัวแปรที่อยู่ในตัวแบบมีจำนวนลดลงและในขณะเดียวกันประสิทธิภาพในการพยากรณ์ในข้อมูลชุดทดสอบมีค่าสูงขึ้น วิธี L0L2Learn นั้นมีความโดดเด่นด้านการคัดเลือกตัวแปรเข้าสู่ตัวแบบในทุกกรณี แต่มีความสามารถด้านการพยากรณ์ไม่สูงนักเมื่อเปรียบเทียบกับวิธีที่ใช้ L_1norm ในการวิเคราะห์อย่างวิธี L1 วิธี A-L1L2 และวิธี A-L1 แต่เมื่อพิจารณาจำนวนตัวแปรที่คัดเลือกเข้าสู่ตัวแบบ พบว่า มีจำนวนตัวแปรอิสระที่ถูกคัดเลือกเข้าสู่ตัวแบบน้อยกว่าวิธี A-L1L2 และวิธี A-L1 และมีค่าใกล้เคียงกับจำนวนตัวแปรอิสระที่มีค่าสัมประสิทธิ์ไม่เท่ากันศูนย์ในความเป็นจริง และสามารถอธิบายได้ง่ายกว่าวิธีอื่น ๆ กล่าวคือตัวแบบที่ได้จากวิธี L0L2Learn มีความโดดเด่นด้านความบางเบาและความสามารถในการอธิบาย ส่วนวิธี L0Learn นั้นพบว่ามีการคัดเลือกตัวแปรได้ไม่ดีนักแต่มีความสามารถในการพยากรณ์เมื่อข้อมูลมีค่า SNR น้อย ในการคัดเลือกตัวแปรเข้าสู่ตัวแบบของวิธี L0Learn นั้น จะมีจำนวนตัวแปรในตัวแบบน้อยกว่าวิธีอื่น ๆ อย่างไรก็ตามในบางครั้งจะไม่มีตัวแปรอิสระใดถูกคัดเลือกเข้าสู่ตัวแบบเลยทำให้ตัวแบบที่ได้ไม่สามารถอธิบายได้

สรุปผลการวิเคราะห์ ในด้านการพยากรณ์พบว่า เมื่อข้อมูลมีค่า SNR ต่ำและตัวแปรอิสระมีความสัมพันธ์กันน้อยถึงปานกลาง วิธี L1 จะมีประสิทธิภาพสูงที่สุด ตามด้วยวิธี L0L2Leran วิธี L0Learn วิธี A-L1L2 และวิธี A-L1 ตามลำดับ แต่เมื่อข้อมูลมีค่า SNR เพิ่มสูงขึ้นและในขณะเดียวกันตัวแปรอิสระมีความสัมพันธ์กันมากขึ้นวิธี A-L1 และวิธี A-L1L2 จะมีประสิทธิภาพสูงที่สุด ตามด้วยวิธี L1 วิธี L0L2Leran วิธี L0Learn ตามลำดับ ส่วนผลการเปรียบเทียบประสิทธิภาพด้านการคัดเลือกตัวแปรเข้าสู่ตัวแบบ เมื่อพิจารณาจากค่าเฉลี่ยของค่า Precision วิธี L0Learn และวิธี L0L2Learn มีประสิทธิภาพมากกว่าวิธีอื่น ๆ และเมื่อพิจารณาจากค่าเฉลี่ยของค่า Recall ในกรณีข้อมูลมีค่า SNR ต่ำวิธี A-L1 และวิธี A-L1L2 จะมีประสิทธิภาพมากที่สุด รองลงมาคือวิธี L0L2Learn วิธี L1 และวิธี L0Learn ตามลำดับ แต่เมื่อข้อมูลมีค่า SNR มากขึ้นและตัวแปรอิสระมีความสัมพันธ์กันมากขึ้น วิธี L1 มีประสิทธิภาพสูงที่สุดเทียบเท่ากับวิธี A-L1 และวิธี A-L1L2 และเมื่อพิจารณาจากค่าเฉลี่ยของค่า AUC กรณีข้อมูลมีค่า SNR ต่ำและตัวแปรอิสระมีความสัมพันธ์กันน้อย วิธี L0L2Learn วิธี L1 วิธี A-L1 และวิธี A-L1L2 จะมีประสิทธิภาพใกล้เคียงกันและมีประสิทธิภาพมากกว่าวิธี L0Learn แต่เมื่อข้อมูลมีค่า SNR มากขึ้นและตัวแปรอิสระมีความสัมพันธ์กันมากขึ้นวิธี L0L2Learn และวิธี A-L1L2 จะมีประสิทธิภาพดีกว่าวิธีอื่น ๆ นอกจากนี้ยังพบว่าวิธี L0Learn และวิธี L0L2Learn จะให้ตัวแบบที่มีขนาดเล็กส่งผลให้ตัวแบบมีค่า Precision โดยเฉลี่ยสูง และมีข้อดีคือตัวแบบอธิบายได้ง่าย ในทางตรงกันข้ามวิธี L1 วิธี A-L1 และวิธี A-L1L2 จะให้ตัวแบบที่มีขนาดใหญ่กว่าส่งผลให้มีค่า Recall โดยเฉลี่ยสูง แต่มีข้อจำกัดคือตัวแบบอธิบายได้ยาก

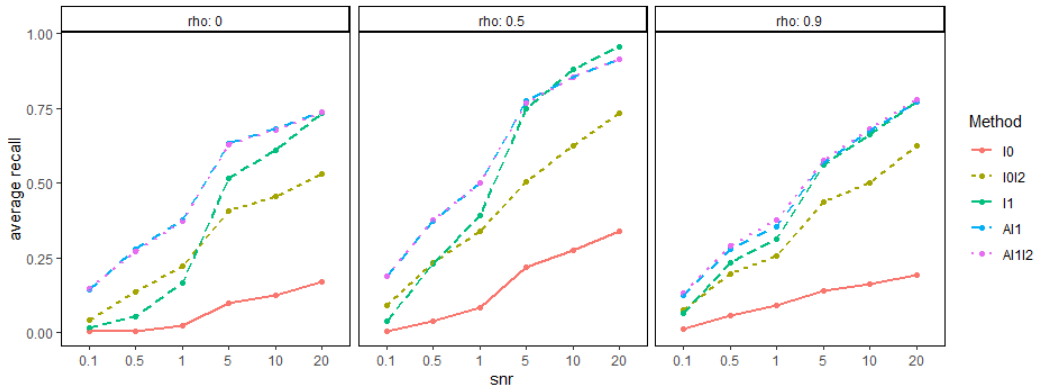


รูปที่ 5.1 ค่า Log ของค่าเฉลี่ยค่าความคลาดเคลื่อนกำลังสอง (log(AMSE)) เมื่อข้อมูลมีค่า SNR และตัวแปรอิสระมีความสัมพันธ์กันในระดับแตกต่างกัน

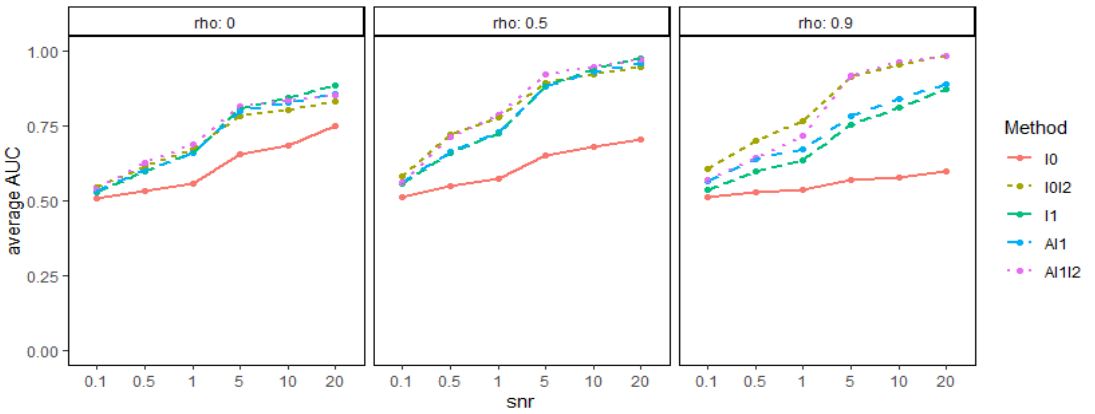


รูปที่ 5.2 ค่าเฉลี่ยของค่า Precision เมื่อข้อมูลมีค่า SNR และตัวแปรอิสระมีความสัมพันธ์กันในระดับแตกต่างกัน

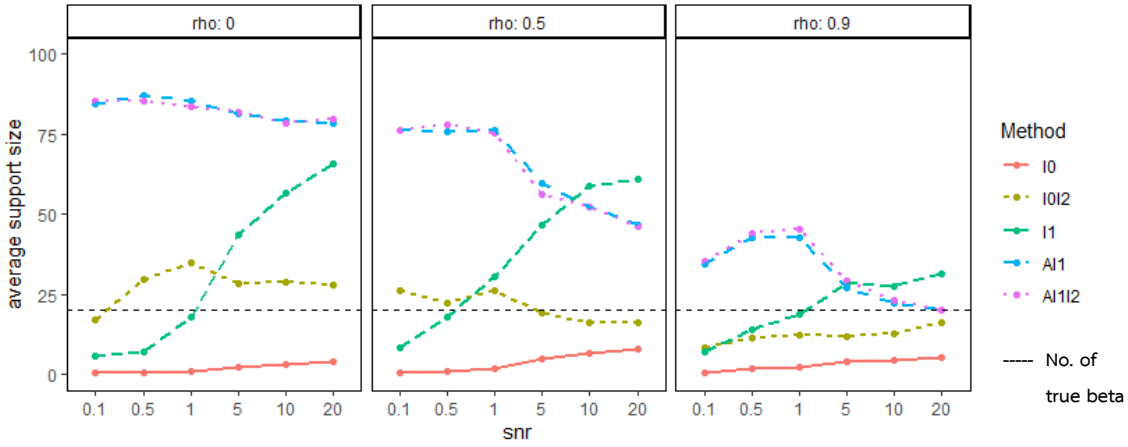
จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University



รูปที่ 5.3 ค่าเฉลี่ยของค่า Recall เมื่อข้อมูลมีค่า SNR และตัวแปรอิสระมีความสัมพันธ์กันในระดับแตกต่างกัน



รูปที่ 5.4 ค่าเฉลี่ยของค่าพื้นที่ใต้กราฟ ROC (Average of AUC Score) เมื่อข้อมูลมีค่า SNR และตัวแปรอิสระมีความสัมพันธ์กันในระดับแตกต่างกัน



รูปที่ 5.5 ค่าเฉลี่ยของจำนวนตัวแปรอิสระที่ถูกคัดเลือกเข้าสู่ตัวแบบ เมื่อข้อมูลมีค่า SNR และตัวแปรอิสระมีความสัมพันธ์กันในระดับแตกต่างกัน

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY

5.2 อภิปรายผลการวิจัย

จากการศึกษาพฤติกรรมของวิธีวิเคราะห์การถดถอยสำหรับข้อมูลที่มีมิติสูงสังเกตได้ว่าวิธี L0Learn มีค่า Precision สูงมากและสูงกว่าทุก ๆ กรณี เนื่องจากการวิธี L0Learn มีการคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบค่อนข้างน้อย และน้อยกว่าจำนวนตัวแปรอิสระที่มีค่าสัมประสิทธิ์ไม่เท่ากับศูนย์ที่แท้จริง โดยเฉพาะกรณีที่ข้อมูลมีตัวแปรอิสระมีความสัมพันธ์กันอย่างสุดโต่ง คือมีความสัมพันธ์น้อยมาก หรือมีความสัมพันธ์มากเกินไป และข้อมูลมีสัญญาณรบกวนสูง วิธี L0Learn จะไม่เลือกตัวแปรอิสระใดเข้าสู่ตัวแบบเลย สอดคล้องกับการศึกษาของ Hazimeh และ Mazumder ในปี 2020 และเหตุที่เป็นเช่นนี้เนื่องมาจากการวิเคราะห์ขาด Function ที่เข้ามาช่วยในการหดตัวของค่าสัมประสิทธิ์ของตัวแปร (Hazimeh & Mazumder, 2020; Mazumder et al., 2017) นอกจากนี้วิธี

L0Learn ยังมีประสิทธิภาพในการพยากรณ์ต่ำมาก เมื่อเปรียบเทียบกับวิธีอื่น ๆ ทั้งนี้เป็นเพราะวิธี L0Learn เกิด Overfitting ได้ง่าย ตัวแบบจะยึดติดกับข้อมูลชุดทดลองมากเกินไป ทำให้ค่า MSE สูง (Mazumder et al., 2017) อย่างไรก็ตามกลับพบว่าผลการทดลองมีความไม่สอดคล้องกับการศึกษาของ Zhang และ Zhang (Zhang & Zhang, 2012) โดยพบว่าตัวแบบที่ได้จากการวิเคราะห์มีความไม่คงเส้นคงวานัก สังเกตได้จากค่าส่วนเบี่ยงเบนมาตรฐานของค่า Precision ค่า Recall และค่า AUC Score ที่ค่อนข้างสูง (จากตารางที่ 4.4-4.12 ในบทที่ 4)

ส่วนวิธี L0L2Learn ซึ่งถูกพัฒนาเพื่อแก้ปัญหา Overfitting ของวิธี L0Learn แล้ว พบว่าสามารถลดปัญหาดังกล่าวได้จริง โดยพบว่าค่า AMSE ลดลงเป็นอย่างมาก ตัวแปรที่ถูกคัดเลือกเข้าสู่ตัวแบบมีจำนวนมากขึ้น และใกล้เคียงกับจำนวนตัวแปรที่มีสัมประสิทธิ์ไม่เป็นศูนย์ที่แท้จริง อีกทั้งยังให้สามารถคัดเลือกตัวแปรได้แม่นยำมากขึ้น นอกจากนี้เมื่อเปรียบเทียบกับวิธี A-L1L2 ซึ่งได้รับการพัฒนาจากจุดเด่นของวิธี L1 และ A-L1 (Zou & Zhang, 2009) จนอาจจะถือว่าเป็นตัวแบบที่ดีที่สุดในตระกูล $L_1 norm$ พบว่าในด้านความสามารถและประสิทธิภาพในการคัดเลือกตัวแปรเข้าสู่ตัวแบบนั้นวิธี L0L2Learn มีประสิทธิภาพสูงทัดเทียมกับวิธี A-L1L2 และมีประสิทธิภาพสูงกว่าในกรณีที่ข้อมูลมีค่า SNR ต่ำ แต่ในกรณีที่ข้อมูลมีค่า SNR สูงวิธี A-L1L2 มีข้อได้เปรียบเพราะวิธี A-L1L2 มีการวิเคราะห์แบบ Conservative โดยผสมผสานทั้งเรื่องความไม่แน่นอนของตัวแบบ การคัดเลือกตัวแปรเข้าสู่ตัวแบบ การประมาณค่าสัมประสิทธิ์การถดถอยและความแปรปรวน ทำให้วิธี A-L1L2 มีประสิทธิภาพทั้งด้านการพยากรณ์และการคัดเลือกตัวแปร (Morozova et al., 2015) อย่างไรก็ตามหากพิจารณาความบางเบาของตัวแบบแล้ววิธี L0L2Learn จะมีจำนวนตัวแปรที่ถูกคัดเลือกเข้าสู่ตัวแบบใกล้เคียงกับจำนวนตัวแปรอิสระที่มีค่าสัมประสิทธิ์ไม่เท่ากับศูนย์ที่แท้จริงมากที่สุด ทำให้ตัวแบบที่ได้มีทั้งความบางเบาและยังสามารถอธิบายตัวแบบได้ง่าย ส่วนวิธี A-L1L2 แม้จะมีความสามารถในการพยากรณ์แต่จำนวนตัวแปรอิสระที่อยู่ในตัวแบบมีจำนวนมากเกินไป โดยเฉพาะกรณีที่ข้อมูลมีตัวแปรอิสระที่ไม่มีความสัมพันธ์กัน และมีค่า SNR ต่ำ เมื่อเปรียบเทียบกับวิธี L0L2Learn แล้ว พบว่าวิธี A-L1L2 จะคัดเลือกตัวแปรเข้าสู่ตัวแบบมากกว่าวิธี L0L2Learn อยู่ประมาณ 2-4 เท่าตัว

วิธี L0L2Learn จะความโดดเด่นด้านความบางเบา ตัวแบบอธิบายได้ง่าย แต่มีความแม่นยำในการพยากรณ์น้อยกว่า ส่วนวิธี A-L1L2 มีความโดดเด่นด้านความสามารถในการพยากรณ์ แต่ตัวแบบไม่บางเบา อธิบายได้ยาก ซึ่งในบางครั้งผู้วิเคราะห์จะต้องเลือก (Trade-off) ระหว่างตัวแบบที่มีความสามารถด้านการทำนายหรือตัวแบบที่อธิบายได้ แต่ก็มีนักวิจัยหลายคนมีความเห็นว่าตัวแบบที่สามารถอธิบายได้มีความสำคัญมากกว่าตัวแบบที่ทำนายได้แม่นยำ ความสามารถในการอธิบายตัวแบบจะไม่สำคัญก็ต่อเมื่อการวิจัยอยู่บานพื้นฐานของสถานการณ์มีความเสี่ยงต่ำ คือ ความผิดพลาดจะไม่ส่งผลกระทบร้ายแรง หรือสถานการณ์ดังกล่าวถูกศึกษาจนเข้าใจอย่าละเอียดถี่ถ้วนแล้ว (Molnar, 2020) แต่เมื่อใดก็ตามที่ปัญหายังไม่ได้รับการศึกษาจนเข้าใจอย่างถ่องแท้ ตัวแบบที่ใช้จะต้องเน้นตัว

แบบที่สามารถอธิบายได้ (Interpretable) คือสามารถอธิบายได้ว่าตัวแปรใดหรือปัจจัยใดที่ส่งผลต่อการพยากรณ์ (Doshi-Velez & Kim, 2017) เพราะการทำนายที่แม่นยำเป็นเพียงส่วนหนึ่งของการทำความเข้าใจปัญหาเท่านั้น ด้วยสาเหตุนี้ความสามารถในการอธิบายจึงมีความสำคัญมากกว่าความสามารถในการพยากรณ์ (Doshi-Velez & Kim, 2017; Miller, 2019; Molnar, 2020)

5.3 ข้อจำกัดของงานวิจัย และข้อเสนอแนะ

งานวิจัยครั้งนี้เป็นการศึกษาครั้งนี้เป็นการศึกษาเชิงทดลอง ซึ่งข้อมูลที่ใช้เป็นข้อมูลที่ได้จากการจำลองทั้งหมด อีกทั้งยังมีการสร้างสถานการณ์ซึ่งอาจเป็นสถานการณ์สุดโต่ง เช่น ข้อมูลมีตัวแปรอิสระที่ไม่มีความสัมพันธ์กันเลย หรือตัวแปรอิสระมีความสัมพันธ์กันสูงมากถึง 0.9 ซึ่งในความเป็นจริงแล้วอาจไม่พบข้อมูลลักษณะนี้มากนัก อีกทั้งงานวิจัยนี้ยังขาดการทดลองกับชุดข้อมูลจริง ทำให้ไม่สามารถเปรียบเทียบให้เห็นภาพได้ว่าเมื่อใช้การวิเคราะห์กับชุดข้อมูลจริงแล้ว ทั้ง 5 วิธีจะมีพฤติกรรมและประสิทธิภาพคล้ายกับชุดข้อมูลจำลองมากน้อยเพียงใด นอกจากนี้งานวิจัยครั้งนี้ศึกษาเพียงข้อมูลที่มีมิติสูงที่มีการแจกแจงแบบปกติหลายตัวแปร (Multivariate Normal Distribution) เท่านั้น ยังไม่ได้ครอบคลุมถึงข้อมูลที่มีการแจกแจงในลักษณะอื่น ๆ รวมไปถึงยังขาดการเปรียบเทียบกับการวิเคราะห์ข้อมูลมิติสูงวิธีอื่น ๆ เช่น วิธี Stepwise วิธี Ridge Regression วิธี Elastic Net Regression วิธี Empirical Bayes Variable Selection เป็นต้น และนอกจากสามารถวิเคราะห์ข้อมูลที่มีมิติสูงแล้ว วิธี L0Learn L0L2Learn L1 A-L1 และ A-L1L2 ยังสามารถวิเคราะห์ข้อมูลมิติต่ำหรือข้อมูลที่มีจำนวนตัวอย่างมากกว่าจำนวนตัวแปรอิสระได้ด้วย แต่ในการศึกษาครั้งนี้ยังไม่ได้ศึกษา ทำให้การศึกษายังไม่ครอบคลุมถึงสถานการณ์ที่ข้อมูลมีลักษณะเป็นข้อมูลมิติต่ำ

จากข้อจำกัดของงานวิจัยในครั้งนี้ที่ได้กล่าวมาจึงเกิดเป็นช่องว่างในงานวิจัย ซึ่งสามารถในไปพัฒนางานวิจัยในอนาคตได้ ดังนี้

1) ควรมีการวิเคราะห์ในชุดข้อมูลจริง เพื่อให้เห็นพฤติกรรมของวิธีต่าง ๆ ที่เกิดขึ้น และสามารถเปรียบเทียบผลลัพธ์ระหว่างข้อมูลที่สร้างขึ้นกับชุดข้อมูลจริงได้ จะทำให้เห็นภาพการเปรียบเทียบที่ชัดเจนมากขึ้น

2) ศึกษาข้อมูลที่มีลักษณะการแจกแจงอื่น ๆ นอกเหนือจากการแจกแจงแบบปกติหลายตัวแปร

3) อาจใช้วิธีการวิเคราะห์อื่น ๆ ที่น่าสนใจ เช่น วิธี Ridge Regression วิธี Elastic net Regression วิธี Empirical Bayes Variable Selection เป็นต้น เพื่อทำการเปรียบเทียบผลลัพธ์ของการศึกษาทั้งด้านความแม่นยำในการพยากรณ์ และความสามารถในการคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบ

4) ศึกษาในชุดข้อมูลที่มีจำนวนตัวอย่างมากกว่าจำนวนตัวแปรอิสระ เพื่อเปรียบเทียบพฤติกรรมของวิธีวิเคราะห์ทั้ง 5 วิธี

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

บรรณานุกรม

Baynes, R. E., Dix, K. J., & Riviere, J. E. (2012). Distribution and pharmacokinetics models. In *Pesticide Biotransformation and Disposition* (pp. 117-147). Academic Press.

Bertsimas, D., King, A., & Mazumder, R. (2016). Best subset selection via a modern optimization lens. *The Annals of Statistics*, *44*(2), 813-852. https://doi.org/10.1214/15-aos1388

Breiman, L. (1992). The little bootstrap and other methods for dimensionality selection in regression: X-fixed prediction error. *Journal of the American Statistical Association*, *87*(419), 738-754.

Bühlmann, P., & Van De Geer, S. (2011). *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media.

Chen, S., Donoho, D., & Saunders, M. (1998). Atomic decomposition by basis pursuit: SIAM Journal on Scientific Computing.

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

Fan, J., & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, *96*(456), 1348-1360.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, *33*(1), 1.

Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction* (Vol. 2). Springer.

Hastie, T., Tibshirani, R., & Tibshirani, R. J. (2017). Extended comparisons of best subset selection, forward stepwise selection, and the lasso. *arXiv preprint arXiv:1707.08692*.

Hastie, T., Tibshirani, R., & Wainwright, M. (2015). Statistical learning with sparsity. *Monographs on statistics and applied probability*, *143*, 143.

Hazimeh, H., & Mazumder, R. (2020). Fast Best Subset Selection: Coordinate Descent and Local Combinatorial Optimization Algorithms. *Operations Research.*, *68(5)*,

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

1517-1537. https://doi.org/10.1287/opre.2019.1919

Hazimeh, H., Mazumder, R., & Nonet, T. (2022). L0Learn: A Scalable Package for Sparse Learning using L0 Regularization. *arXiv preprint arXiv:2202.04820*.

Huang, J., Ma, S., & Zhang, C.-H. (2008). Adaptive Lasso for sparse high-dimensional regression models. *Statistica Sinica*, 1603-1618.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning with Applications in R*. Springer Science+Business Media. https://doi.org/10.1007/978-1-4614-7138-7

Mazumder, R. (2020). Discussion of "Best Subset, Forward Stepwise or Lasso? Analysis and Recommendations Based on Extensive Comparisons". *Statistical Science*, *35*(4), 602-608. https://doi.org/10.1214/20-sts807

Mazumder, R., Radchenko, P., & Dedieu, A. (2017). Subset selection with shrinkage: Sparse linear modeling when the SNR is low. *arXiv preprint arXiv:1708.03288*.

Meinshausen, N., & Bühlmann, P. (2006). Variable selection and high-dimensional graphs with the lasso. *Ann Stat*, *34*, 1436-1462.

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, *267*, 1-38.

Molnar, C. (2020). *Interpretable machine learning*. Lulu. com.

Morozova, O., Levina, O., Uusküla, A., & Heimer, R. (2015). Comparison of subset selection methods in linear regression in the context of health-related quality of life and substance abuse in Russia. *BMC medical research methodology*, *15*(1), 1-17.

Pungpapong, V., Zhang, M., & Zhang, D. (2015). Selecting massive variables using an iterated conditional modes/medians algorithm. *Electronic Journal of Statistics*, *9*(1), 1243-1266.

Pungpapong, V., Zhang, M., & Zhang, D. (2015). Selecting massive variables using an iterated conditional modes/medians algorithm. *Electronic Journal of Statistics*, *9*, 1243-1266.

Raskutti, G., Wainwright, M. J., & Yu, B. (2011). Minimax rates of estimation for high-dimensional linear regression over $\ell_q $-balls. *IEEE transactions on information theory*, *57*(10), 6976-6994.

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

Syed, A. R. (2011). A review of cross validation and adaptive model selection.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, *58*(1), 267-288.

Trevor Hastie, Robert Tibshirani, & Friedman, J. (2017). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2 ed.). Springer. https://doi.org/10.1007/b94608

Van't Veer, L. J., Dai, H., Van De Vijver, M. J., He, Y. D., Hart, A. A., Mao, M., Peterse, H. L., Van Der Kooy, K., Marton, M. J., & Witteveen, A. T. (2002). Gene expression profiling predicts clinical outcome of breast cancer. *nature*, *415*(6871), 530-536.

Zhang, C.-H., & Zhang, T. (2012). A general theory of concave regularization for high-dimensional sparse estimation problems. *Statistical Science*, *27*(4), 576-593.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association*, *101*(476), 1418-1429.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, *67*(2), 301-320. https://doi.org/10.1111/j.1467-9868.2005.00527.x

Zou, H., & Zhang, H. H. (2009). On the Adaptive Elastic-Net with a Diverging Number of Parameters. *Ann Stat*, *37*(4), 1733-1751. https://doi.org/10.1214/08-AOS625

เบญจมาส รุ่งศรานนท์, & อัชฌา อระวีพร. (2019). การเปรียบเทียบวิธีการประมาณค่าพารามิเตอร์ของการวิเคราะห์การถดถอยที่ปรับด้วยฟังก์ชันการลงโทษภายใต้ข้อมูลที่มีมิติสูง. วารสารวิทยาศาสตร์และเทคโนโลยี, *28*(8 สิงหาคม 2563), 1346-1358. https://doi.org/10.14456/tstj.2020.108

ชุติกาญจน์ ชูสวัสดิ์. (2560). การเปรียบเทียบประสิทธิภาพของวิธีการวิเคราะห์การถดถอยแบบพีนอลไลซ์ ในตัวแบบการถดถอยปัวซง ภายใต้ข้อมูลที่มีมิติสูงแบบบางเบา และตัวแปร อิสระมีความสัมพันธ์กันสูง มหาวิทยาลัยธรรมศาสตร์].

พัชราภรณ์ พรดำเนินสวัสดิ์. (2560). การเปรียบเทียบประสิทธิภาพของการประมาณค่าพารามิเตอร์ด้วยวิธีแลซโซในการถดถอยเชิงเส้นที่มีมิติสูง มหาวิทยาลัยศิลปากร].

วิฐรา พึ่งพาพงศ์. (2558). บทวิเคราะห์วิธีวิเคราะห์การถดถอยเชิงเส้นสำหรับข้อมูลที่มีมิติสูง. วารสารวิทยาศาสตร์และเทคโนโลยี., *23(2)*, 212-223.

สวรรยา ภู่เงิน. (2557). การเปรียบเทียบวิธีคัดกรองตัวแปรสำหรับการทดสอบกลุ่มของสัมประสิทธิ์การ

ถดถอยที่มีมิติสูงแบบลำดับชั้น [วิฐรา พึ่งพาพงศ์, จุฬาลงกรณ์มหาวิทยาลัย].

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University

ประวัติผู้เขียน

ชื่อ-สกุล	วรัญญา บุตรบุรี
วัน เดือน ปี เกิด	14 สิงหาคม 2536
สถานที่เกิด	จังหวัดนครพนม
วุฒิการศึกษา	เศรษฐศาสตร์บัณฑิต มหาวิทยาลัยเชียงใหม่ (เกียรตินิยมอันดับหนึ่ง)
ที่อยู่ปัจจุบัน	63 หมู่ 11 ตำบลก้านเหลือง อำเภอนาแก จังหวัดนครพนม 48130

จุฬาลงกรณ์มหาวิทยาลัย
CHULALONGKORN UNIVERSITY