การเปรียบเทียบวิธีการคัดกรองตัวแปรสำหรับวิธีการแบ่งข้อมูลตัวอย่างหลายครั้งในการหาค่าพี-แวลูสำหรับข้อมูลที่มีมิติสูง

นายศุภวัฒน์ อังคะสี

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
สาขาวิชาสถิติ ภาควิชาสถิติ
คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2556

COMPARISON OF THE VARIABLES SCREENING METHODS FOR MULTI – SAMPLE SPLIT TO FIND P – VALUES FOR HIGH – DIMENSIONAL DATA

Mr. Supawat Angkasi

A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Science Program in Statistics
Department of Statistics
Faculty of Commerce and Accountancy
Chulalongkorn University
Academic Year 2013

หัวข้อวิทยานิพนธ์	การเปรียบเทียบวิธีการคัดกรองตัวแปรสำหรับวิธีการแบ่งข้อมูลตัวอย่างหลายครั้งในการหาค่าพี-แวลูสำหรับข้อมูลที่มีมิติสูง
โดย	นายศุภวัฒน์ อังคะสี
สาขาวิชา	สถิติ
อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก	อาจารย์ ดร. วิฐรา พึ่งพาพงศ์

คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญามหาบัณฑิต

.....คณบดีคณะพาณิชยศาสตร์และการบัญชี
(รองศาสตราจารย์ ดร. พสุ เดชะรินทร์)

คณะกรรมการสอบวิทยานิพนธ์

.....ประธานกรรมการ
(รองศาสตราจารย์ ดร. สุพล ดุรงค์วัฒนา)

.....อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก
(อาจารย์ ดร. วิฐรา พึ่งพาพงศ์)

.....กรรมการ
(อาจารย์ ดร. อัครินทร์ ไพบูลย์พานิช)

.....กรรมการภายนอกมหาวิทยาลัย
(อาจารย์ ดร. อรุณี กำลัง)

ศุภวัฒน์ อังคะสี : การเปรียบเทียบวิธีการคัดกรองตัวแปรสำหรับวิธีการแบ่งข้อมูลตัวอย่างหลายครั้งในการหาค่าพี-แวลูสำหรับข้อมูลที่มีมิติสูง. (COMPARISON OF THE VARIABLES SCREENING METHODS FOR MULTI – SAMPLE SPLIT TO FIND P – VALUES FOR HIGH – DIMENSIONAL DATA) อ.ที่ปรึกษาวิทยานิพนธ์หลัก: อ. ดร. วิฐรา พึ่งพาพงศ์, 104 หน้า.

งานวิจัยฉบับนี้มีวัตถุประสงค์เพื่อเปรียบเทียบวิธีการคัดกรองตัวแปรจากวิธี Lasso, Adaptive Lasso, Elastic net และ SCAD สำหรับขั้นตอนวิธีแบ่งข้อมูลหลายครั้ง (Multi - Split) เพื่อหาค่า p-value ในการวิเคราะห์ความถดถอยของข้อมูลที่มีมิติสูง โดยวิเคราะห์จากจำนวนสัมประสิทธิ์ของตัวแปรอิสระที่ไม่เท่ากับ 0 ความผิดพลาดเชิงบวกและความผิดพลาดเชิงลบภายหลังจากควบคุมด้วยวิธี False Discovery Rate (FDR) โดยมีการจำลองข้อมูลที่มีขอบเขตต่างกัน โดยมีขนาดตัวอย่างเท่ากับ 10, 100 และ 200 จำนวนสัมประสิทธิ์ที่ไม่เท่ากับ 0 เป็นร้อยละ 10, 20, 50 ของขนาดตัวอย่าง และความสัมพันธ์ของตัวแปรอิสระเป็น 0, 0.5 และ 0.9 โดยทำการจำลองข้อมูลและวิเคราะห์ผลด้วยโปรแกรม R 3.0.3 ทั้งนี้จะใช้ค่าความผิดพลาดในการตรวจจับเชิงบวก (False Positive : FP) ความผิดพลาดในการตรวจจับเชิงลบ (False Negative : FN) และจำนวนของสัมประสิทธิ์ของตัวแปรอิสระที่มีค่าไม่เท่ากับ 0 จากการทดสอบสมมติฐานเมื่อควบคุม FDR เป็นเครื่องมือในการเปรียบเทียบและการวัดประสิทธิภาพ

การศึกษาภายใต้ขอบเขตดังกล่าวผลปรากฏว่ากรณีที่ขนาดตัวอย่างเท่ากับ 10 พิจารณาจากจำนวนของสัมประสิทธิ์ของตัวแปรอิสระที่มีค่าไม่เท่ากับ 0 จากการทดสอบสมมติฐาน เมื่อควบคุม FDR ,ค่าของ FP และ FN ที่ตารางแสดงจำนวนของสัมประสิทธิ์ของตัวแปรอิสระที่มีค่าไม่เท่ากับ 0 จากการทดสอบสมมติฐาน เมื่อควบคุม FDR และค่าของ FN จะไปในทิศทางเดียวกัน นั่นคือการคัดกรองตัวแปรด้วยวิธี Adaptive Lasso จะเหมาะสมมากที่สุด แต่จากตาราง FP จะได้วิธี Lasso ที่เหมาะสมแต่ค่าที่ได้ยังไม่ชัดเจน ในกรณีที่ขนาดตัวอย่างเท่ากับ 100 และ 200 การคัดกรองตัวแปรด้วยวิธี Adaptive Lasso และวิธี SCAD จะเหมาะสมมากที่สุด แต่จากตาราง FP จะได้วิธี Lasso และวิธี EN ที่เหมาะสม นั่นแสดงให้เห็นว่าวิธี Lasso และวิธี EN มีประสิทธิภาพในการคัดกรองตัวแปรน้อยกว่าวิธี Adaptive Lasso และวิธี SCAD

ภาควิชา สถิติ ลายมือชื่อนิสิต .....

สาขาวิชา สถิติ ลายมือชื่อ อ.ที่ปรึกษาวิทยานิพนธ์หลัก .....

ปีการศึกษา 2556

5581609426 : MAJOR STATISTICS
KEYWORDS: HIGH DIMENSIONAL DATA / FALSE DISCOVERY RATE / MILTI – SPLIT

SUPAWAT ANGKASI: COMPARISON OF THE VARIABLES SCREENING METHODS FOR MULTI – SAMPLE SPLIT TO FIND P – VALUES FOR HIGH – DIMENSIONAL DATA. ADVISOR: VITARA PUNGPAPONG, Ph.D., 104 pp.

This research is aimed to compare the screening variables of Lasso, Adaptive Lasso, Elastic net and SCAD for the Multi - Split to find p-values in the regression analysis for high dimensional data. To analyze from the number of non-zero coefficients, false positives and false negatives after controlling False Discovery Rate (FDR) were collected and analyzed based on simulated data. The sample size are 10, 100 and 200. The numbers of non-zero coefficients is not equal to 0 are set to 10, 20 and 50 percent of sample size and the correlation among independent variables are 0, 0.5 and 0.9. he simulating and analyzing data in this study used the R 3.0.3 . It uses The False Positive (FP), The False Negative (FN) and the number of coefficients of independent variables is not equal to 0 by hypothesis testing after control by FDR., which is not equal to 0, that use as a tool to compare and performance measurement.

The study showed that within the scope of the case considering the sample size of 10 .The tables of the number of coefficients of independent variables is not equal to 0 by hypothesis testing after control by FDR, FP and FN shows the value of the number of coefficients of independent variables is not equal to 0 by hypothesis testing after control by FDR and FN are go to the same direction. That is data screening by Adaptive Lasso are the most appropriate. On the other hand, in the table of FP data screening by Lasso, this will get to the right value but the value will not very clear. In case of the sample size are 100 and 200, the data screening by Adaptive Lasso and SCAD are the most appropriate but from the table of FP will approach Lasso and appropriate EN, which showed that Lasso and EN are effective to the data screening,that is less than Adaptive Lasso and SCAD.

Department: Statistics Student's Signature .....

Field of Study: Statistics Advisor's Signature .....

Academic Year: 2013

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้เสร็จสมบูรณ์ลงได้ด้วยดี ด้วยความช่วยเหลือและความเอาใจใส่จาก อาจารย์ ดร.วิฐรา พึ่งพาพงศ์ อาจารย์ที่ปรึกษาวิทยานิพนธ์ ผู้วิจัยขอกราบขอบพระคุณท่านอาจารย์เป็นอย่างสูง ที่กรุณาให้คำปรึกษา อบรมสั่งสอน และให้ข้อคิดเห็นต่างๆ ตลอดจนความช่วยเหลือคำแนะนำเพื่อปรับปรุงแก้ไขวิทยานิพนธ์ และเป็นกำลังใจในการทำงาน จนกระทั่งวิทยานิพนธ์เสร็จสมบูรณ์ด้วยดี

ผู้วิจัยขอกราบขอบพระคุณท่าน รองศาสตราจารย์ ดร. สุพล ดุรงค์วัฒนา ประธานกรรมการสอบวิทยานิพนธ์ อาจารย์ ดร.อัครินทร์ ไพบูลย์พานิช และอาจารย์ ดร.อรุณี กำลัง กรรมการสอบวิทยานิพนธ์เป็นอย่างสูงที่ท่านอาจารย์ทั้งสามท่านได้เสียสละเวลาเพื่อสอบ ตรวจสอบและให้คำแนะนำเพื่อแก้ไขวิทยานิพนธ์ฉบับนี้ให้สมบูรณ์ยิ่งขึ้น อีกทั้งขอกราบขอบพระคุณคณาจารย์ประจำภาควิชาสถิติ คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัยทุกท่านที่ให้โอกาสทางการศึกษา และอบรมสั่งสอนความรู้ทั้งในการเรียนและการดำรงชีวิตให้แก่ผู้วิจัยจนกระทั่งเสมอมาจนสำเร็จการศึกษาในครั้งนี้

สุดท้ายนี้ผู้วิจัยขอกราบขอบพระคุณครอบครัว ที่ให้กำลังและให้ความห่วงใย ส่งเสริมและสนับสนุนมาโดยตลอด และขอขอบคุณเพื่อน ๆ ทุกคน ที่คอยช่วยเหลือ ให้คำแนะนำและเป็นกำลังใจให้กับผู้วิจัยตลอดมา

สารบัญ

หน้า

หน้า

สารบัญตาราง

ตารางที่ หน้า

ตารางที่ หน้า

สารบัญภาพ

ภาพที่ หน้า

บทที่ 1

บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

เป็นที่น่าสังเกตว่าข้อมูลที่ใช้ในการวิเคราะห์ทางสถิติในปัจจุบันไม่ว่าจะเป็นในด้านของการวิจัยทางการแพทย์หรือทางการเงินการธนาคารล้วนแล้วแต่มีฐานข้อมูลขนาดใหญ่ เนื่องจากความเจริญก้าวหน้าทางด้านเทคโนโลยีและต้นทุนที่ใช้ในการเก็บข้อมูลต่ำลง จึงสามารถเก็บข้อมูลได้รวดเร็วและในปริมาณที่มาก ดังนั้นจึงมีผู้ที่สนใจที่จะวิเคราะห์เพื่อศึกษาข้อมูลเหล่านี้เพิ่มมากขึ้น และวิธีวิเคราะห์ข้อมูลที่แพร่หลายวิธีหนึ่งก็คือการวิเคราะห์การถดถอย ซึ่งใช้ในการวิเคราะห์ความสัมพันธ์ของตัวแปรอิสระ (Independent Variables) และตัวแปรตาม (Dependent Variables) นอกจากนี้ตัวแบบที่ได้ยังสามารถนำไปใช้ในการพยากรณ์ได้ อย่างไรก็ตามหากตัวแปรอิสระ(p) มีจำนวนมากกว่าขนาดตัวอย่าง(n) การประมาณค่าสัมประสิทธิ์ในตัวแบบการวิเคราะห์การถดถอยโดยวิธีกำลังสองน้อยที่สุด(Ordinary Least Squares (OLS)) จะหาค่าไม่ได้ ข้อมูลลักษณะดังกล่าวเราเรียกว่าข้อมูลที่มีมิติสูง (High-Dimensional Data) ปัจจุบันมีการเสนอวิธีการเพื่อใช้ในการหาตัวแบบการถดถอยที่เหมาะสมสำหรับข้อมูลที่มีมิติสูง เช่น วิธี Penalized Likelihood Estimator โดยเป็นการเพิ่ม Likelihood ด้วย Penalty Function ซึ่งเป็นฟังก์ชันในรูปของสัมประสิทธิ์และหาค่าสัมประสิทธิ์ที่ทำให้ค่า Penalized Likelihood ดังกล่าวมีค่าสูงสุด หากเราเลือกใช้ Penalty Function ที่เหมาะสม จะทำให้ค่าสัมประสิทธิ์บางตัวมีค่าเท่ากับศูนย์ โดยที่วิธี Penalized Likelihood Estimator จะทำการประมาณค่าสัมประสิทธิ์ไปพร้อมๆกับการคัดเลือกตัวแปรเข้ามายังตัวแบบ

วิธี Least Absolute Shrinkage and Selection Operator (Lasso) นำเสนอโดย Tibshirani (1996) ในวิธี Lasso จะมีการใช้ l_1 – norm สำหรับ Penalty Function ในการปรับค่าผลรวมความคลาดเคลื่อนกำลังสองให้มีค่าน้อยที่สุด วิธี Lasso มีข้อดีคือจะทำการประมาณสัมประสิทธิ์แบบต่อเนื่อง (Continuous Shrinkage) ไปพร้อมๆกับการคัดเลือกตัวแปรเข้ามายังตัวแบบ แต่วิธี Lasso ยังมีข้อเสียคือ ตัวประมาณค่าสัมประสิทธิ์ที่ได้มีความเอนเอียง (Bias)

วิธี Adaptive Least Absolute Shrinkage and Selection Operator (Adaptive Lasso) ซึ่งนำเสนอโดย Zou (2006)เป็นวิธีที่พัฒนามาจากวิธี Lasso ซึ่งมีการเพิ่มเงื่อนไขโดยการให้ค่าน้ำหนัก (Weight) ให้กับพารามิเตอร์แต่ละตัวแตกต่างกันใน Penalty Function ซึ่งการเพิ่มค่าน้ำหนักนี้จะช่วยแก้ไขปัญหาความไม่คงเส้นคงวาที่ทำให้เกิดความเอนเอียงในวิธี Lasso ได้อีกด้วย ซึ่ง

การใช้ Penalty Function ดังกล่าวจะทำให้ได้ตัวประมาณที่มีคุณสมบัติ Oracle ของตัวประมาณค่า กล่าวคือ เมื่อขนาดตัวอย่างเข้าสู่ค่าอนันต์ วิธี Adaptive Lasso จะมีความสามารถในการเลือกตัวแปรได้เสมือนกับว่าทราบตัวแบบที่แท้จริง (True Model) ซึ่งคุณสมบัตินี้ไม่มีในวิธี Lasso

วิธี Elastic Net (EN) นำเสนอโดย Zou and Hastie (2003) มีคุณสมบัติช่วยลดค่าสัมประสิทธิ์ของตัวแปรแบบต่อเนื่องและทำการคัดเลือกตัวแปรไปพร้อมๆกัน โดยที่ EN มีประสิทธิภาพที่ดีกว่าวิธี Lasso ในกรณีที่ตัวพยากรณ์ในแต่ละตัวแบบมีความสัมพันธ์กันมาก โดยเฉพาะอย่างยิ่งในกรณีที่จำนวนตัวแปรอิสระมีขนาดใหญ่กว่าขนาดของตัวอย่างมากๆ โดยวิธี EN มีการให้ Penalty Function อยู่ในรูปของ
l_1 – norm และ l_2 – norm

วิธี The Smoothly Clipped Absolute Deviation (SCAD) นำเสนอโดย Fan and Li (2001) มีคุณสมบัติเพื่อเป็นการลดค่าเอนเอียง (Bias) สำหรับค่าประมาณสัมประสิทธิ์ของตัวแปรอิสระที่ได้ โดยที่ Penalty Function จะแบ่งออกเป็น 3 กรณีด้วยกัน ซึ่งจะแบ่งตามขอบเขตของ $|\beta_j|$ ทำให้ได้ตัวประมาณที่มีคุณสมบัติ Oracle ของตัวประมาณค่าเหมือนวิธี Adaptive Lasso โดยเมื่อขนาดตัวอย่างเข้าสู่ค่าอนันต์ วิธี SCAD จะมีความสามารถในการเลือกตัวแปรได้เสมือนกับว่าทราบตัวแบบที่แท้จริง (True Model)

อย่างไรก็ตามวิธีประมาณค่าและคัดกรองตัวแปรที่ได้กล่าวมาทั้งหมดนี้จะทำให้ค่าประมาณของสัมประสิทธิ์การถดถอยที่ได้ส่วนใหญ่มีค่าเท่ากับ 0 แต่จะไม่มีการรายงานค่า p-value ซึ่งใช้วัดระดับความสำคัญของตัวแปรอิสระ ดังนั้นการคัดกรองตัวแปร (Variable Screening) ในขั้นตอนแรกจึงอาศัยเพียงค่าสัมประสิทธิ์การถดถอยของตัวแปรเท่านั้น และในปี 2008 Wasserman และ Roeder ได้เสนอวิธีการในการหาค่า p - value สำหรับข้อมูลที่มีมิติสูง โดยวิธีการแบ่งข้อมูลออกเป็น 2 ส่วนเท่าๆกัน โดยข้อมูลส่วนแรกใช้ในการคัดกรองตัวแปร(Variable Screening) ซึ่งสามารถควบคุมขนาดของตัวแปรอิสระได้และส่วนที่สองใช้ในการหาค่า p - value ด้วยวิธี OLS สำหรับการแบ่งข้อมูลนั้นจะทำโดยสุ่มและเพื่อการควบคุมให้ผลลัพธ์ที่ได้ไม่เกิดความเอนเอียง (Bias) Meinshausen, Meier et al. (*2009)* จึงได้เสนอวิธีที่ทำการแบ่งข้อมูลออกเป็น *2* ส่วนโดยสุ่มหลายๆครั้ง (Multi - Split) โดยการนำข้อมูลส่วนที่สองที่ได้จากการแบ่งในแต่ละครั้งมาหาค่า p-value ของตัวประมาณสัมประสิทธิ์การถดถอยทั้งหมดด้วยวิธี OLS แล้วนำค่า p-value ที่ได้ในแต่ละครั้งมาทำการปรับค่าโดยยังไม่มีการรวมค่า หลังจากนั้นจึงนำค่า p-value ที่ทำการปรับค่าแล้วทั้งหมดมาทำการรวมค่าอีกครั้งด้วยฟังก์ชันควอนไทล์ (Quantile Function)

เมื่อได้ค่า p-value สำหรับการทดสอบสัมประสิทธิ์การถดถอยจากวิธี Multi-Split เป็นที่เรียบร้อยแล้วจึงนำค่า p-value ดังกล่าวมาใช้ในการคัดเลือกตัวแปรในขั้นตอนสุดท้าย โดยใช้การ

ควบคุม False Discovery Rate (FDR) ซึ่งนำเสนอโดย Benjamini และ Hochberg ในปี 1993 เพื่อใช้ในการควบคุมอัตราส่วนของความผิดพลาดที่เกิดขึ้นท่ามกลางกลุ่มของการปฏิเสธสมมติฐานว่าง โดยกำหนดให้ Q แทนสัดส่วนของจำนวนความผิดพลาดชนิดที่ 1 (Type I Error) หรือ False Discoveries ภายใต้จำนวนการปฏิเสธสมมติฐานว่าง (H_0) ทั้งหมดเพื่อลดความผิดพลาดในการคัดเลือกตัวแปรเข้ามายังตัวแบบได้อีกด้วย ในงานวิจัยนี้ไม่เลือกใช้วิธี Family Wise Error Rate (FWER) ซึ่งเป็นวิธีการควบคุมความผิดพลาดที่ไม่เหมาะสมที่จะนำมาใช่เนื่องจากไม่สามารถปฏิเสธสมมติฐานว่าง (Null Hypothesis) ที่ว่าสัมประสิทธิ์ของตัวแปรแต่ละตัวมีค่าเท่ากับ 0 เพื่อใช้ในการคัดเลือกตัวแปรได้ แต่เลือกใช้วิธี FDR เพื่อควบคุมความผิดพลาดที่เกิดขึ้นโดยวิธี FDR สามารถที่จะคัดเลือกตัวแปรอิสระได้จากการปฏิเสธสมมติฐานว่างดังกล่าว ดังนั้นจึงทำให้วิธี FDR มีประสิทธิภาพมากกว่าและยังมีเหมาะสมที่จะใช้ในกรณีที่มีจำนวนตัวแปรอิสระจำนวนมากอีกด้วย

ในการศึกษาครั้งนี้ ผู้วิจัยสนใจในการหาวิธีการคัดกรองตัวแปรที่ดีที่สุดในขั้นแรกของวิธี Multi-Split จากวิธี Lasso, Adaptive Lasso, EN, และ SCAD แล้วนำตัวแปรที่ได้จากการคัดกรองจากทั้ง 4 วิธีที่กล่าวมานั้นมาใช้ในการหาค่า p-value และนำค่า p-value ที่ได้นี้มาใช้ในการคัดเลือกตัวแปรขั้นสุดท้ายโดยการควบคุม FDR

1.2 วัตถุประสงค์

เพื่อศึกษาเปรียบเทียบวิธีการคัดกรองตัวแปรสำหรับขั้นตอนวิธี Multi – Split เพื่อหาค่า p-value ในการวิเคราะห์ความถดถอยของข้อมูลที่มีมิติสูง

1.3 ข้อตกลงเบื้องต้น

ศึกษาตัวแปรภายใต้รูปแบบความสัมพันธ์การถดถอยเชิงเส้น ในรูป

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon \tag{1.1}$$

จาก $X_i \sim N(0,\Sigma)$, $\varepsilon \sim N(0,\sigma^2 I_n)$ เมื่อ $\sigma^2 = 1$

โดยที่ $\Sigma = \begin{bmatrix} \rho_{11} & \cdots & \rho_{1p} \\ \vdots & \ddots & \vdots \\ \rho_{p1} & \cdots & \rho_{pp} \end{bmatrix}$; $\rho_{ij} = \begin{cases} 1 & ; \ i = j \\ \rho^{|i-j|} & ; \ i \neq j \end{cases}$

$\mathbf{Y} = (y_1, y_2, \ldots, y_n)'$ เป็นเวกเตอร์ของตัวแปรตามขนาด n x1

$\mathbf{X} = (x_1, x_2, \ldots, x_n)'$ เป็นเมตริกซ์ตัวแปรอิสระขนาด n x p โดยที่ $X_i = \begin{bmatrix} x_{i1} \\ \vdots \\ x_{ip} \end{bmatrix}$

$\beta = (\beta_1, \beta_2, \ldots, \beta_p)'$ เป็นเวกเตอร์ค่าสัมประสิทธิ์การถดถอยขนาด p x 1

$\varepsilon = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)'$ เป็นเวกเตอร์ความคลาดเคลื่อนขนาด n x 1

เมื่อ n คือขนาดตัวอย่าง

p คือจำนวนตัวแปรอิสระ

1.4 ขอบเขตของการวิจัย

ในการศึกษานี้จะทำการศึกษาภายใต้ขอบเขตดังนี้

1. ศึกษาภายใต้อัตราส่วนของขนาดตัวอย่างต่อจำนวนตัวแปรอิสระ (n : p) ที่ 1:2, 1:5 และ 1:10

   กรณีที่ 1 : n = 10 เปรียบเทียบอัตราส่วน n:p ที่ 10 : 20, 10 : 50, 10 : 100

   กรณีที่ 2 : n = 100 เปรียบเทียบอัตราส่วน n:p ที่ 100 : 200, 100 : 500, 100 : 1,000

   กรณีที่ 3 : n = 200 เปรียบเทียบอัตราส่วน n:p ที่ 200 : 400, 200 : 1,000, 200 : 2,000
2. ศึกษาภายใต้ขนาด (Effect Size) ของค่าสัมประสิทธิ์ตัวที่ไม่เท่ากับ 0 โดยแบ่งเป็นขนาดเล็ก $(0 < |\beta| < 1)$ และขนาดใหญ่ $(1 < |\beta| < 10)$
3. ศึกษาภายใต้ร้อยละของจำนวนสัมประสิทธิ์ (β) ที่ไม่เท่ากับ 0 เมื่อเทียบกับขนาดตัวอย่างที่ 10, 20 และ 50
4. ศึกษาภายใต้ความสัมพันธ์ (Correlation) ของตัวแปรอิสระ 3 ระดับ โดยมีจำนวนตัวแปรอิสระ(p) ที่ใช้ในงานวิจัยทั้งหมด 8 ระดับ คือ 20, 50, 100, 200, 400, 500, 1000, 2000 ตัว จะมีเมตริกซ์ความแปรปรวนร่วมของแต่ละระดับความสัมพันธ์ ดังนี้

เมตริกซ์ความแปรปรวนร่วม $(\boldsymbol{\Sigma}) = \begin{bmatrix} \rho_{11} & \rho_{12} & \cdots & \rho_{1p} \\ \rho_{21} & \rho_{22} & \cdots & \rho_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{p1} & \rho_{p2} & \cdots & \rho_{pp} \end{bmatrix}_{pxp}$

โดยที่ $\rho_{ij} = \begin{cases} 1 & ; \ i = j \\ \rho^{|i-j|} & ; \ i \neq j \end{cases}$

โดยความสัมพันธ์ (Correlation) ของตัวแปรอิสระทั้ง 3 ระดับ คือ

ระดับที่ 1 : ρ = 0

ระดับที่ 2 : ρ = 0.5

ระดับที่ 3 : ρ = 0.9

<u>กรณีที่มี $\rho = 0$</u>

ที่จำนวนตัวแปรอิสระทั้งหมด 8 ระดับ คือ 20, 50, 100, 200, 400, 500, 1000, 2000 ตัว

จะได้ว่า $\rho_{ij} = 0$ เท่ากันในทุกกรณี

<u>กรณีที่มี $\rho = 0.5$</u>

ที่จำนวนตัวแปรอิสระเท่ากับ 20 ค่า จะได้ว่า $\rho_{ij} \in [1.9\times10^{-6},0.5]$; $i \neq j$

ที่จำนวนตัวแปรอิสระเท่ากับ 50 ค่า จะได้ว่า $\rho_{ij} \in [1.7\times10^{-15},0.5]$; $i \neq j$

ที่จำนวนตัวแปรอิสระเท่ากับ 100 ค่า จะได้ว่า $\rho_{ij} \in [1.6\times10^{-30},0.5]$; $i \neq j$

ที่จำนวนตัวแปรอิสระเท่ากับ 200 ค่า จะได้ว่า $\rho_{ij} \in [1.2\times10^{-60},0.5]$; $i \neq j$

ที่จำนวนตัวแปรอิสระเท่ากับ 400 ค่า จะได้ว่า $\rho_{ij} \in [0,0.5]$; $i \neq j$

ที่จำนวนตัวแปรอิสระเท่ากับ 500 ค่า จะได้ว่า $\rho_{ij} \in [0,0.5]$; $i \neq j$

ที่จำนวนตัวแปรอิสระเท่ากับ 1000 ค่า จะได้ว่า $\rho_{ij} \in [0,0.5]$; $i \neq j$

ที่จำนวนตัวแปรอิสระเท่ากับ 2000 ค่า จะได้ว่า $\rho_{ij} \in [0,0.5]$; $i \neq j$

<u>กรณีที่มี $\rho = 0.9$</u>

ที่จำนวนตัวแปรอิสระเท่ากับ 20 ค่า จะได้ว่า $\rho_{ij} \in [0.14,0.9]$; $i \neq j$

ที่จำนวนตัวแปรอิสระเท่ากับ 50 ค่า จะได้ว่า $\rho_{ij} \in [0.0057,0.9]$; $i \neq j$

ที่จำนวนตัวแปรอิสระเท่ากับ 100 ค่า จะได้ว่า $\rho_{ij} \in [0.000029,0.9]$; $i \neq j$

ที่จำนวนตัวแปรอิสระเท่ากับ 200 ค่า จะได้ว่า $\rho_{ij} \in [7.8\times10^{-10},0.9]$; $i \neq j$

ที่จำนวนตัวแปรอิสระเท่ากับ 400 ค่า จะได้ว่า $\rho_{ij} \in [5.5\times10^{19},0.9]$; $i \neq j$

ที่จำนวนตัวแปรอิสระเท่ากับ 500 ค่า จะได้ว่า $\rho_{ij} \in [1.4\times10^{-23},0.9]$; $i \neq j$

ที่จำนวนตัวแปรอิสระเท่ากับ 1000 ค่า จะได้ว่า $\rho_{ij} \in [1.9\times10^{-46},0.9]$; $i \neq j$

ที่จำนวนตัวแปรอิสระเท่ากับ 2000 ค่า จะได้ว่า $\rho_{ij} \in [3.4\times10^{-92},0.9]$; $i \neq j$

1.5 คำจำกัดความที่ใช้ในงานวิจัย

ข้อมูลที่มีมิติสูง (High-Dimensional Data)

คือ ข้อมูลที่มีจำนวนตัวแปรอิสระ(p) มากกว่าขนาดตัวอย่าง(n)

การแบ่งข้อมูลโดยสุ่มหลายๆครั้ง (The Multi - Split)

คือ การแบ่งข้อมูลออกเป็น 2 ส่วนโดยสุ่มหลายๆครั้ง

อัตราส่วนความผิดพลาดที่เกิดขึ้น (False Discovery Rate : FDR)

คือ อัตราส่วนความผิดพลาดที่เกิดขึ้นท่ามกลางกลุ่มของการปฏิเสธสมมติฐานว่าง

ความผิดพลาดในการตรวจจับเชิงบวก (False Positive : FP)

คือ การวัดจำนวนที่เกิดความผิดพลาดจากข้อสรุปที่ค่าประมาณสัมประสิทธิ์การถดถอยเชิงเส้นมีค่าไม่เท่ากับศูนย์ เมื่อค่าสัมประสิทธิ์การถดถอยที่แท้จริงมีค่าเท่ากับศูนย์

ความผิดพลาดในการตรวจจับเชิงลบ (False Negative : FN)

คือ การวัดจำนวนที่เกิดความผิดพลาดจากข้อสรุปที่ค่าประมาณสัมประสิทธิ์การถดถอยเชิงเส้นมีค่าเท่ากับศูนย์ เมื่อค่าสัมประสิทธิ์การถดถอยที่แท้จริงมีค่าไม่เท่ากับศูนย์

1.6 เกณฑ์ที่ใช้ในการตัดสินใจ

เกณฑ์ที่ใช้ในการตัดสินใจว่าวิธีการคัดกรองตัวแปรวิธีการใดเหมาะสมในการคัดกรองตัวแปรสำหรับขั้นตอนวิธี Multi –Split เพื่อหาค่า p-value ในการวิเคราะห์การถดถอยของข้อมูลที่มีมิติสูงมากที่สุด โดยจะพิจารณาจาก ค่าความผิดพลาดในการตรวจจับเชิงบวก (FP) ความผิดพลาดในการตรวจจับเชิงลบ (FN) และจำนวนของสัมประสิทธิ์ของตัวแปรอิสระที่มีค่าไม่เท่ากับ 0 จากการทดสอบสมมติฐาน เมื่อควบคุม FDR ดังนี้

กำหนดให้ $S = \{ j : \beta_j \neq 0 \}$

$\hat{S} = \{ j :$ ปฏิเสธ $H_0 : \beta_j = 0 \}$

1. ความผิดพลาดในการตรวจจับเชิงบวก (False Positive : FP)

คือ การวัดจำนวนที่เกิดความผิดพลาดจากข้อสรุปที่ค่าประมาณสัมประสิทธิ์การถดถอยเชิงเส้นมีค่าไม่เท่ากับศูนย์ เมื่อค่าสัมประสิทธิ์การถดถอยที่แท้จริงมีค่าเท่ากับศูนย์สามารถคำนวณได้ดังนี้

$$FP = |\hat{S} \cap S^c| \tag{1.2}$$

2. ความผิดพลาดในการตรวจจับเชิงลบ (False Negative : FN)

คือ การวัดจำนวนที่เกิดความผิดพลาดจากข้อสรุปที่ค่าประมาณสัมประสิทธิ์การถดถอยเชิงเส้นมีค่าเท่ากับศูนย์ เมื่อค่าสัมประสิทธิ์การถดถอยที่แท้จริงมีค่าไม่เท่ากับศูนย์ สามารถคำนวณได้ดังนี้

$$FN = |\hat{S}^c \cap S| \tag{1.3}$$

3. จำนวนของสัมประสิทธิ์ของตัวแปรอิสระที่มีค่าไม่เท่ากับ 0 จากการทดสอบสมมติฐาน เมื่อควบคุม FDR หรือ $|\hat{S}|$

โดยถ้าวิธีใดให้ค่าวัดประสิทธิภาพ FP และ FN ต่ำที่สุด และเซตของตัวแปรที่มีค่าสัมประสิทธิ์ไม่เท่ากับ 0 มีขนาดใกล้เคียงกับตัวแบบที่แท้จริงมากที่สุดจะถือได้ว่าวิธีนั้นมีความเหมาะสมที่จะใช้ใน

การพิจารณาคัดกรองตัวแปรและมีความเหมาะสมที่จะใช้ในการประมาณค่าสัมประสิทธิ์ของตัวแปรอิสระได้อีกด้วย

1.7 วิธีการศึกษา

1. ศึกษาตัวแบบและทฤษฎีที่เกี่ยวข้อง
2. กำหนดและจำลองข้อมูล
   2.1 กำหนดค่าเริ่มต้นโดยการสร้างข้อมูลที่มีจำนวนค่าสังเกต n ค่า และจำนวนพารามิเตอร์ p ตัว โดยใช้อัตราส่วน n:p คือ
   - 10:20, 10:50 และ 10:100
   - 100:200, 100:500 และ 100:1000
   - 200:400, 200:1000 และ 200:2000

   2.2 กำหนดให้ร้อยละของจำนวนสัมประสิทธิ์ (β) ที่ไม่เท่ากับ 0 เมื่อเทียบกับขนาดตัวอย่างที่ 10, 20 และ 50
   2.3 กำหนดให้ขนาด (Effect Size) ของสัมประสิทธิ์ที่ไม่เท่ากับ 0 ในกรณีขนาดเล็ก $(0 < |\beta| < 1)$ และขนาดใหญ่ $(1 < |\beta| < 10)$
   2.4 จำลองข้อมูลที่มีการแจกแจงแบบปกติ (Normal Distribution) ภายใต้รูปแบบความสัมพันธ์การถดถอยเชิงเส้น ในรูป

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon$$

จาก $\mathbf{X}_i \sim \mathrm{N}(0, \mathbf{\Sigma})$, $\varepsilon \sim \mathrm{N}(0, \sigma^2 \boldsymbol{I}_n)$ เมื่อ $\sigma^2 = 1$

โดยที่ $\mathbf{\Sigma} = \begin{bmatrix} \rho_{11} & \cdots & \rho_{1p} \\ \vdots & \ddots & \vdots \\ \rho_{p1} & \cdots & \rho_{pp} \end{bmatrix}$; $\rho_{ij} = \begin{cases} 1 & ; \ i = j \\ \rho^{|i-j|} & ; \ i \neq j \end{cases}$

$\mathbf{Y} = (y_1, y_2, \ldots, y_n)'$ เป็นเวกเตอร์ของตัวแปรตามขนาด n x1

$\mathbf{X} = (x_1, x_2, \ldots, x_n)'$ เป็นเมตริกซ์ตัวแปรอิสระขนาด n x p โดยที่ $\mathbf{X}_i = \begin{bmatrix} x_{i1} \\ \vdots \\ x_{ip} \end{bmatrix}$

$\beta = (\beta_1, \beta_2, \ldots, \beta_p)'$ เป็นเวกเตอร์ค่าสัมประสิทธิ์การถดถอยขนาด p x 1

$\varepsilon = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)'$ เป็นเวกเตอร์ความคลาดเคลื่อนขนาด n x 1

เมื่อ n คือขนาดตัวอย่าง
p คือจำนวนตัวแปรอิสระ

3. นำข้อมูลที่จำลองขึ้นมาศึกษาและใช้วิธี Lasso, Adaptive Lasso, EN, SCAD ในขั้นตอนการคัดกรองข้อมูลสำหรับวิธีการ Multi - Split ในการคำนวณค่า p-value
4. นำค่า p-value ที่ได้จากการประมาณค่าวิธีต่างๆในข้อ 3. มาคัดเลือกตัวแปรในขั้นสุดท้ายด้วยวิธีการควบคุม False Discovery Rate (FDR)
5. นำข้อมูลที่ได้จากข้อ 4 มาหาค่า FP, FN และ $|\hat{S}|$
6. วิเคราะห์ผลลัพธ์

1.8 ประโยชน์ที่คาดว่าจะได้รับ

สามารถหาวิธีการคัดกรองตัวแปรสำหรับ Multi – Split เพื่อหาค่า p-value สำหรับข้อมูลที่มีมิติสูงได้เหมาะสมโดยพิจารณาจากการควบคุม False Discovery Rate (FDR)

บทที่ 2

ทฤษฎีและตัวสถิติที่เกี่ยวข้อง

การประมาณค่าสัมประสิทธิ์การถดถอย (β) ของข้อมูลในตัวแบบเพื่อคัดเลือกตัวแปรเข้ามายังตัวแบบในกรณีที่ข้อมูลมีจำนวนตัวอย่างมากกว่าจำนวนของตัวแปรอิสระ(n>p) สามารถทำได้โดยวิธีกำลังสองน้อยที่สุด (Odinary Least Square : OLS) แต่เนื่องจากปัจจุบันข้อมูลมีขนาดใหญ่ขึ้นดังนั้นจึงมีข้อมูลที่มีจำนวนของแปรอิสระมากกว่าขนาดตัวอย่าง (p > n) เกิดขึ้นและเรียกข้อมูลประเภทนี้ว่า “ข้อมูลที่มีมิติสูง (High-Dimensional Data)” โดยการประมาณค่าสัมประสิทธิ์ของตัวแปรอิสระในกรณีเป็นข้อมูลที่มีมิติสูงนั้นไม่สามารถหาจากวิธี OLS ได้ ดังนั้นในงานวิจัยนี้จะกล่าวถึงวิธีการคัดกรองตัวแปรสำหรับข้อมูลที่มีมิติสูง คือ วิธี Lasso วิธี Adaptive Lasso วิธี EN และวิธี SCAD แต่เนื่องจากทั้ง 4 วิธีขั้นต้นยังไม่มีการให้ค่า p-value ดังนั้นจึงจะกล่าวถึงวิธี Multi – Split ซึ่งเป็นวิธีที่ใช้ในการหาค่า p-value สำหรับข้อมูลที่มีมิติสูงและยังจะกล่าวถึงวิธีการควบคุมความผิดพลาดที่เกิดขึ้นท่ามกลางกลุ่มของการปฏิเสธสมมติฐานว่าง(FDR) และเกณฑ์ที่ใช้ในการตัดสินใจ คือ ความผิดพลาดในการตรวจจับเชิงบวก (False Positive : FP) ความผิดพลาดในการตรวจจับเชิงลบ (False Negative : FN) และจำนวนของสัมประสิทธิ์ของตัวแปรอิสระที่มีค่าไม่เท่ากับ 0 จากการทดสอบสมมติฐาน เมื่อควบคุม FDR

2.1 การประมาณค่าสัมประสิทธิ์การถดถอยด้วยวิธีกำลังสองน้อยที่สุด

เนื่องจากข้อมูลส่วนใหญ่มีขนาดใหญ่ขึ้นและข้อจำกัดภายใต้งบประมาณและเวลาดังนั้นการรวบรวมข้อมูลของประชากรทั้งหมดจึงทำได้ยากทำให้ไม่ทราบค่าพารามิเตอร์ของประชากร เพราะสามารถที่จะรวบรวมข้อมูลได้เฉพาะค่าสังเกตของตัวอย่างเท่านั้น ดังนั้นจึงมีการเสนอวิธีการประมาณค่าสัมประสิทธิ์สำหรับข้อมูลตัวอย่างที่เรียกว่า วิธีกำลังสองน้อยที่สุด (Ordinary Least Squares : OLS)ขึ้นมาเพื่อใช้ในการประมาณค่าสัมประสิทธิ์ของตัวแปรอิสระแต่ละตัว

โดยกำหนดให้ $\hat{\beta}$ เป็นเวกเตอร์ของค่าประมาณของ β

โดยที่ $$\hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \vdots \\ \hat{\beta}_p \end{bmatrix}$$

e เป็นเวกเตอร์ของค่าประมาณความคลาดเคลื่อน ε เรียกเป็น เวกเตอร์ส่วนเหลือ

โดยที่ $$\mathbf{e} = \begin{bmatrix} e_1 \\ \vdots \\ e_p \end{bmatrix} = \begin{bmatrix} Y_1 - \hat{Y}_1 \\ \vdots \\ Y_p - \hat{Y}_p \end{bmatrix}$$

เมื่อ $\hat{Y}_i$, i = 1, 2, ..., n คือค่าประมาณของ Y_i

วิธีการประมาณแบบกำลังสองน้อยที่สุด เป็นวิธีการหาค่า $\hat{\beta}_0$, $\hat{\beta}_1$, ..., $\hat{\beta}_p$ หรือ $\hat{\beta}$ เพื่อใช้เป็นค่าประมาณของ β_0, β_1 , ..., β_p หรือ β ที่ทำให้ผลบวกกำลังสองของความคลาดเคลื่อน $\mathbf{e}'\mathbf{e}$ มีค่าน้อยที่สุด ทำได้โดยการหาอนุพันธ์ของ $\sum e_i^2$ เทียบกับ $\hat{\beta}_0$, $\hat{\beta}_1$, ..., $\hat{\beta}_p$ แล้วกำหนดให้เท่ากับ 0

ให้ $\tilde{\mathbf{X}} = (1_n \mathbf{X})$ โดยที่ 1_n คือเวกเตอร์ที่มีค่า 1 ทั้งหมด ขนาด n

ถ้า $\mathbf{X}'\mathbf{X}$ ไม่เป็นเมตริกซ์เอกฐาน (Nonsingular Matrix) จะได้ตัวประมาณของ β แบบกำลังสองน้อยที่สุด คือ

$$\hat{\beta} = (\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'Y \tag{2.1}$$

มีค่าความแปรปรวน คือ

$$V(\hat{\beta}) = (\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}\sigma^2 \tag{2.2}$$

และเวกเตอร์ส่วนเหลือ คือ

$$\mathbf{e} = Y - \hat{Y} = [I_n - H]Y \tag{2.3}$$

เมตริกซ์ความแปรปรวน – ความแปรปรวนร่วมของ e คือ

$$V(\mathbf{e}) = (I_n - H)\sigma^2 \tag{2.4}$$

เมื่อ I คือ เมตริกซ์เอกลักษณ์ (Identity Matrix) ขนาด n x n

H คือ เมตริกซ์ที่เกิดจากค่าของตัวแปรอิสระเท่านั้น เรียกว่า Hat Matrix

โดยที่ $H = (\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'$

จะเห็นว่าความคลาดเคลื่อน ε_i กับส่วนเหลือ e_i แตกต่างกัน ε_i เป็นตัวแปรสุ่มที่ถูกสมมติว่าไม่มีสหสัมพันธ์กันมีค่าเฉลี่ยเท่ากับ 0 และความแปรปรวนคงที่เท่ากับ σ^2 แต่ส่วนเหลือนั้นมีสหสัมพันธ์กัน ค่าเฉลี่ยเท่ากับ 0 และมีความแปรปรวนที่ไม่คงที่

การทดสอบสมมติฐานเกี่ยวกับพารามิเตอร์ของตัวแบบ

การทดสอบสมมติฐานเกี่ยวกับพารามิเตอร์ β_j เมื่อ j = 1, 2, 3, ..., p เป็นการทดสอบว่า $\beta_j = 0$ หรือไม่ ถ้า $\beta_j = 0$ แล้ว จะหมายความว่าไม่มีความสัมพันธ์เชิงเส้นระหว่างตัวแปรอิสระ X_j และ Y แต่ถ้า $\beta_j \neq 0$ นั่นแปลว่า X_j และ Y มีความสัมพันธ์เชิงเส้น ซึ่งรวมไปถึงว่าการรู้ค่าของ X_j จะช่วยให้การคาดคะเนค่า Y ได้ถูกต้องเพิ่มขึ้น การทดสอบสมมติฐานเกี่ยวกับพารามิเตอร์ β_j จึงเป็นการทดสอบสมมติฐาน

$$H_0 : \beta_j = 0$$

$$H_1 : \beta_j \neq 0$$

โดยทั่วไปค่าพารามิเตอร์ β_0 , β_1 , ..., β_p เราจะไม่สามารถทราบค่าได้ดังนั้นจึงทำการสุ่มตัวอย่างมาจากประชากรแล้วจึงนำมาหาสมการถดถอยเชิงเส้นของกลุ่มตัวอย่าง $\hat{y}_j = \tilde{\beta}_0 + \tilde{\beta}_1 x_{i1} + \ldots + \tilde{\beta}_p x_{ip}$ เราจะใช้ค่าสถิติ $\tilde{\beta}_0$, $\tilde{\beta}_1$, ..., $\tilde{\beta}_p$ ที่ได้จากการสุ่มตัวอย่างไปทดสอบว่าพารามิเตอร์ β_0 , β_1 , ..., β_p ของประชากรว่ามีค่าเท่ากับ 0 หรือไม่

ตัวสถิติที่ใช้ในการทดสอบสมมติฐานดังกล่าวคือ

$$t = \frac{\hat{\beta}_j - \beta_j}{s\sqrt{c_{jj}}} \qquad (2.5)$$

โดยที่ $s = \sqrt{\frac{e'e}{n-p-1}}$ และ c_{jj} คือค่าในตำแหน่งที่ j ในแนวทแยงของเมตริกซ์ $(\tilde{X}'\tilde{X})^{-1}$ เมื่อ H_0 จริง ตัวสถิติทดสอบ t มีการแจกแจงแบบสติวเด้นส์ที มีองศาอิสระ n-p-1 โดยจะปฏิเสธ H_0 ที่ระดับนัยสำคัญ α เมื่อค่า |t| มีค่ามากกว่า $t_{(\frac{\alpha}{2}, n-p-1)}$

การประมาณค่าสัมประสิทธิ์การถดถอยด้วยวิธีกำลังสองน้อยที่สุดสามารถทำงานได้ภายใต้เงื่อนไขที่ขนาดของตัวอย่างใหญ่กว่าจำนวนตัวแปรอิสระ (n > p) แต่ในกรณีที่ขนาดของตัวอย่างที่เล็กกว่าจำนวนตัวแปรอิสระ (n < p) จะไม่สามารถหาค่า $(\tilde{X}'\tilde{X})^{-1}$ ได้เนื่องจาก $\tilde{X}'\tilde{X}$ ไม่เป็นเมตริกซ์เอกฐาน (Nonsingular Matrix) ดังนั้นถ้ากรณีที่ n < p จะสามารถทำการประมาณค่าสัมประสิทธิ์การถดถอยได้ด้วยวิธี Penalized Likelihood

P – Value

ในการทดสอบสมมติฐานจะสนใจว่าสมมติฐานว่างนั้นจะถูกปฏิเสธหรือไม่ หากสมมติฐานว่างไม่ถูกปฏิเสธนั่นไม่ได้หมายความว่าสมมติฐานว่างนั้นเป็นจริงเสมอ แต่ยังไม่มีหลักฐานที่เพียงพอที่จะปฏิเสธสมมติฐานว่าง ในการตัดสินใจว่าจะปฏิเสธสมมติฐานว่างหรือไม่นั้นเราจำเป็นต้องกำหนดระดับนัยสำคัญ โดยที่ α = P(ปฏิเสธ H_0 | H_0 เป็นจริง) การทดสอบสมมติฐานนั้นสามารถทำได้สองวิธีคือ การเปรียบเทียบขอบเขตวิกฤต และ การพิจารณาค่า P-Value

P-Value คือ ความน่าจะเป็นที่ค่าสถิติทดสอบมีค่าเกินกว่าสถิติทดสอบที่คำนวณได้จากข้อมูลตัวอย่าง ภายใต้สมติฐานว่างเป็นจริง ในการคำนวณค่า P-Value อาศัยการ Integration

เพื่อที่จะหาพื้นที่ใต้กราฟ (ในทางสถิติ พื้นที่ใต้กราฟคือความน่าจะเป็น) ในการทดสอบสมมติฐานจะปฏิเสธสมมติฐานว่างก็ต่อเมื่อค่า P-Value < α

การคำนวณ P-Value

กรณีการทดสอบ t-test

สมมติให้ตัวสถิติทดสอบคือ t_{cal} ซึ่งมีการแจกแจง $t_{(n-p-1)}$

หลักการพิจารณาการทดสอบสมมติฐานหางเดียวด้านขวาสามารถคำนวณ P-value ได้ดังนี้

P-value = $P(t_{(n-p-1)} > t_{cal}) = \int_{t_{cal}}^{\infty} f(t)\,dt$

หลักการพิจารณาการทดสอบสมมติฐานหางเดียวด้านซ้ายสามารถคำนวณ P-value ได้ดังนี้

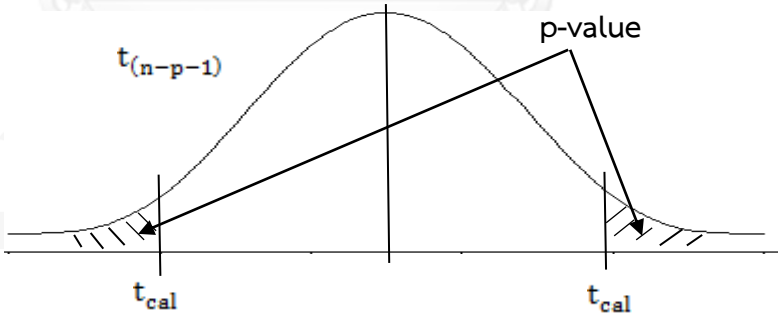
P-value = $P(t_{(n-p-1)} < t_{cal}) = \int_{-\infty}^{t_{cal}} f(t)\,dt$

ในที่นี้ สมมติฐานที่เราต้องการทดสอบคือ $H_0: \beta_j = 0$ และ $H_1: \beta_j \neq 0$ ซึ่งเป็นการทดสอบสมมติฐานแบบสองหาง (Two-tailed test) ดังนั้น

p-value = $P(t_{(n-p-1)} > |t_{cal}|) = \int_{-\infty}^{t_{cal}} f(t)\,dt + \int_{t_{cal}}^{\infty} f(t)\,dt$ ดังรูปที่ 2.1

เมื่อ f(t) คือฟังก์ชันการแจกแจงที (Student's t-distribution)

รูปที่ 2.1 แสดงพื้นที่ของการปฏิเสธสมมติฐานว่างในกรณีที่เป็นการทดสอบสองทางที่มีการแจกแจงแบบปกติ



เมื่อ t_{cal} คือ ค่าสถิติ t ที่ได้จากการคำนวณจากสูตรที่ 2.5

2.2 การประมาณค่าสัมประสิทธิ์การถดถอยด้วยวิธี Penalized Likelihood

การหาค่าสัมประสิทธิ์การถดถอย ที่ทำให้ Penalized Likelihood มีค่าสูงสุด นั่นคือ

$$\hat{\beta} = \arg\min_{\beta} \left(-l(\beta) + P_{\lambda}(\beta)\right), \ \lambda \geq 0 \qquad (2.6)$$

โดย $-l(\beta)$ = - loglikelihood

$P_{\lambda}(\beta)$ คือ Penalty Function

λ คือ Tuning Parameter โดยที่ $\lambda \geq 0$

สมการ (2.6) จะถือเป็นการคัดกรองตัวแปรเข้าในตัวแบบ หากเราเลือกใช้ Penalty Function ที่เหมาะสมจะสามารถคัดกรองตัวแปรเข้าในตัวแบบได้ กล่าวคือ Penalty Function จะทำให้ค่าสัมประสิทธิ์บางตัวมีค่าเท่ากับ 0

2.2.1 Penalty Function ของวิธี Least Absolute Shrinkage and Selection Operator (Lasso)

Tibshirani (1996) ได้เสนอวิธี Lasso โดยใช้ l_1 – norm ในการปรับค่าด้วยวิธีกำลังสองน้อยที่สุดสำหรับ Penalty Function $(P_{\lambda}(\beta))$

$$P_{\lambda}(\beta) = \lambda \sum_{j=1}^{p} |\beta_j| \qquad \text{โดยที่} \qquad \lambda > 0 \qquad (2.7)$$

ซึ่งจากวิธีดังกล่าว Lasso ยังมีข้อเสียคือ การประมาณค่าสัมประสิทธิ์ที่ได้มีความเอนเอียง (Bias)

2.2.2 Penalty Function ของวิธี Adaptive Least Absolute Shrinkage and Selection Operator (Adaptive Lasso)

Zou (2006) ได้เสนอวิธี Adaptive Lasso โดยพัฒนามาจากวิธี Lasso โดยมีการเพิ่มเงื่อนไขโดยการให้ค่าน้ำหนัก (Weight) ให้กับพารามิเตอร์แต่ละตัวแตกต่างกัน แล้วจึงประมาณค่าสัมประสิทธิ์โดยการสร้าง Penalty Function$(P_{\lambda}(\beta))$

$$P_{\lambda}(\beta) = \lambda \sum_{i=1}^{p} \hat{w}_j |\beta_j| \qquad (2.8)$$

โดยที่ $\hat{W}_j \begin{cases} \frac{1}{|\hat{\beta}_{OLS}|} & ; n > p \\ \frac{1}{|\hat{\beta}_{Ridge}|} & ; n < p \end{cases}$

โดยใน Penalty Function นี้ได้มีการกล่าวถึงคุณสมบัติ Oracle ของตัวประมาณค่าจากวิธี Adaptive Lasso ซึ่งคุณสมบัตินี้ไม่มีใน Lasso นั่นคือ เมื่อขนาดตัวอย่างเข้าสู่ค่าอนันต์ Adaptive Lasso จะมีความสามารถในการเลือกตัวแปรได้เสมือนกับว่าทราบตัวแบบที่แท้จริง (True Model)

2.2.3 Penalty Function ของวิธี Elastic Net (EN)

Zou และ Hastie (2003) ได้เสนอวิธี EN ในการคัดกรองตัวแปรเข้าในตัวแบบ ซึ่งมีคุณสมบัติช่วยลดค่าสัมประสิทธิ์ของตัวแปรที่สัมพันธ์กัน แล้วจึงประมาณค่าสัมประสิทธิ์โดยการสร้าง Penalty Function$(P_\lambda(\beta))$

$$P_{\lambda_1,\lambda_2}(\beta) = \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=1}^{p} \beta_j^2 \qquad \text{โดยที่} \qquad \lambda_1 > 0 \text{ และ } \lambda_2 > 0 \qquad (2.9)$$

2.2.4 Penalty Function ของวิธี The Smoothly Clipped Absolute Deviation (SCAD)

เพื่อเป็นการลดค่าเอนเอียง (Bias) สำหรับค่าประมาณสัมประสิทธิ์ของตัวแปรอิสระที่ได้ Fan และ Li (2001) จึงได้เสนอวิธี SCAD ในการสร้าง Penalty Function$(P_\lambda(\beta))$

$$P_\lambda(\beta) = \sum_{i=1}^{p} P_{\lambda j(\beta_j;a)} \qquad (2.10)$$

โดยที่

$$P_{\lambda_j}(\beta) = \begin{cases} \lambda|\beta_j| & ; \quad |\beta_j| < \lambda \\ -(\beta_j^2 - 2a\lambda|\beta_j| + \lambda^2)/\ [2(a-1)] & ; \ \lambda < |\beta_j| < a\lambda, \\ (a+1)\lambda^2/2 & ; \quad |\beta_j| > a\lambda \end{cases} \qquad (2.11)$$

เมื่อ $\lambda > 0$ และ $a > 2$ ซึ่ง Fan และ Li (2001) ได้มีการเสนอให้ใช้ค่า $a = 3.7$

2.3 การหาค่า p-value ของสัมประสิทธิ์ความถดถอยกรณีข้อมูลที่มีมิติสูงโดยวิธี Multi – Split

Meinshausen, et al. (2009) ได้เสนอวิธีการ Multi - Split เพื่อใช้ในการหาค่า p-value ของสัมประสิทธิ์การถดถอยกรณีที่ข้อมูลมีมิติสูง (n < p) โดย Multi – Split มีขั้นตอนดังนี้

เมื่อ b = 1, ..., B

กำหนดให้ $\tilde{S}^{(b)}$ แทนเซตของดรรชนีค่าสัมประสิทธิ์ตัวที่ j ที่ไม่เท่ากับ 0 $\{ j ; \hat{\beta}_j \neq 0 \}$; $j \in \{1,2,3,\dots,p\}$

1. ทำการแบ่งข้อมูลเริ่มต้นโดยวิธีการสุ่มออกเป็น 2 กลุ่มที่มีจำนวนที่เท่ากัน $\left(\frac{n}{2}\right)$ ซึ่งจะได้กลุ่มของ $D_{in}^{(b)}$ และ $D_{out}^{(b)}$
2. ใช้เฉพาะชุดข้อมูล $D_{in}^{(b)}$ ในการประมาณเซต $\tilde{S}^{(b)}$ โดยที่ $\tilde{S}^{(b)}$ คือเซตของดรรชนีค่าสัมประสิทธิ์การถดถอยที่ไม่เท่ากับ 0
3. คำนวณค่า p-value ดังนี้

(a) ใช้เฉพาะชุดข้อมูล $\mathbf{D}_{\mathbf{out}}^{(b)}$ เพื่อการคัดกรองตัวแปรที่เหมาะสมในเซต $\tilde{S}^{(b)}$ โดยพิจารณาจากค่า p – value ตัวที่ j ($\tilde{P}_j^{(b)}$) สำหรับค่า $j \in \tilde{S}^{(b)}$ ที่คำนวณได้จากวิธี Least Square (OLS)

(b) กำหนดค่า p – value ให้เท่ากับ 1 สำหรับ $j \notin \tilde{S}^{(b)}$ $(\tilde{P}_j^{(b)} = 1)$

4. นิยามการปรับค่า p – value (ไม่มีการรวมค่า) คือ ให้ $|\tilde{S}^{(b)}|$ คือจำนวนสมาชิกในเซต $\tilde{S}^{(b)}$

$$P_j^{(b)} = \min\left(\tilde{P}_j^{(b)}|\tilde{S}^{(b)}|, 1\right), j = 1, \dots, p \tag{2.12}$$

โดยที่ตัวแปร 1 ตัวจะได้ค่า p-value ทั้งหมด B ค่า ดังนั้น จะต้องมีการรวมค่า p – value $P_j^{(b)}$ ทั้งหมด B ค่า ซึ่งจะกล่าวต่อไป

ในการรวมค่า p - value ทั้ง B ค่าในแต่ละตัวพยากรณ์ j = 1, … , p Meinhausen, et al. (2009) แนะนำให้ใช้ควอนไทล์ (Quantiles) ดังนี้

$$Q_j(\gamma) = \min\left\{1, q_\gamma\left(\frac{P_j^{(b)}}{\gamma}; b = 1, \dots, B\right)\right\} \tag{2.13}$$

โดย $q_\gamma(.)$ คือฟังก์ชันควอนไทล์ (γ -Quantile Function) และ $\gamma \in (0,1)$

ค่า p-value สำหรับค่าพยากรณ์แต่ละ j เมื่อ j = 1, ..., p มีค่าเท่ากับ $Q_j(\gamma)$ โดยที่ $0 < \gamma < 1$

อาจจะเป็นการยากที่เราจะเลือกค่า γ ที่เหมาะสมเนื่องจากการควบคุมค่าความคลาดเคลื่อนไม่สามารถบอกได้ว่าค่า γ เป็นค่าที่ดีที่สุด ดังนั้นจึงมีการเสนอให้ใช้ตัวที่ปรับค่าได้แทนการเลือกค่าที่เหมาะสม

โดยให้ $\gamma_{min} \epsilon (0,1)$ เป็นขอบเขตล่างสำหรับ γ = 0.05 และกำหนด

$$P_j = \min\{1, (1 - \log\gamma_{min}) \inf_{\gamma \epsilon (\gamma_{min}, 1)} Q_j(\gamma)\} \tag{2.14}$$

2.4 การควบคุม False Discovery Rate (FDR)

การทดสอบสมมติฐานเพื่อหาความสัมพันธ์หรือหาสาเหตุ หรือทดสอบเพื่อตรวจสอบว่าสิ่งที่ผู้วิจัยคาดไว้เป็นจริงหรือไม่นั้น การทดสอบอาจจะเกิดความผิดพลาดในการสรุปเกิดขึ้น โดยเฉพาะอย่างยิ่งถ้าข้อมูลตัวอย่างที่นำมาใช้ในการทดสอบไม่มีคุณภาพเพียงพอ โดยความผิดพลาดที่เกิดขึ้นในการทดสอบแบ่งเป็น 2 ชนิด คือ

1. ความผิดพลาดประเภทที่ 1 (Type I Error)

เป็นความผิดพลาดที่เกิดขึ้นเนื่องจากผู้วิจัยสรุปว่าสมมติฐานว่างไม่จริง (ปฏิเสธ H_0) ทั้งที่ในความเป็นจริงนั้นสมมติฐาน H_0 จริง

α = P (ปฏิเสธ H_0 | H_0 เป็นจริง)

หรือ α = โอกาสที่ผู้วิจัยจะสรุปผิด คือ สรุปว่า H_0 ไม่เป็นจริงทั้งที่ความเป็นจริงแล้วสมมติฐาน H_0 จริงและเรียกค่า α ว่าระดับนัยสำคัญ (Level of Significance)

2. ความผิดพลาดประเภทที่ 2 (Type II Error)

เป็นความผิดพลาดที่เกิดขึ้นจากการที่ผู้วิจัยยอมรับว่า H_0 จริงโดยที่ในความเป็นจริงนั้น H_0 ไม่จริง

β = P (ไม่สามารถปฏิเสธ H_0 | H_0 ไม่จริง)

หรือ β = โอกาสที่ผู้วิจัยจะสรุปผิดโดยสรุปว่า H_0 จริงทั้งที่ความเป็นจริงแล้ว H_0 ไม่จริง

ตารางที่ 2.1

ความผิดพลาดประเภทที่ 1 (Type I Error)และความผิดพลาดประเภทที่ 2 (Type II Error)

สรุปผลการทดสอบ	ความเป็นจริงของประชากร	
	H_0 เป็นจริง	H_0 เป็นเท็จ
ไม่ปฏิเสธ H_0	1-α	Type II Error (β)
ปฏิเสธ H_0	Type I Error (α)	1- β

โดยที่ α = P(Type I Error) และ β = P(Type II Error)

False Discovery Rate (FDR) ถูกเสนอโดย Benjamini และ Hochberg (1993) เป็นขั้นตอนที่ได้จากการเปรียบเทียบกับความผิดพลาดแบบที่ 1และ 2 (Type I and II Error) ดังนี้

ตารางที่ 2.2

False Discovery Rate (FDR)

	สมมติฐานว่างเป็นจริง (H_0)	สมมติฐานทางเลือกเป็นจริง (H_1)	ผลรวม
ไม่มีนัยสำคัญ	U	T	m – R
มีนัยสำคัญ	V	S	R
ผลรวม	m_0	m- m_0	m

โดย m คือ จำนวนการทดสอบสมมติฐานทั้งหมด

m_0 คือ จำนวนของสมมติฐานว่าง (H_0) ที่ไม่สามารถปฏิเสธได้

m-m_0 คือ จำนวนของสมมติฐานว่าง (H_0) ที่ถูกปฏิเสธ

V คือ “False Discoveries : FD”

S คือ จำนวนของ “True Discoveries”

T คือ จำนวนของความผิดพลาดชนิดที่ 2 (Type II Error)

U คือ จำนวนของ True Negatives

R คือ จำนวนของการปฏิเสธสมมติฐานว่าง (H_0) หรือ “Discoveries”

ซึ่ง *“ False Discovery Rate (FDR) ถูกออกแบบมาเพื่อใช้ในการควบคุมอัตราส่วนของความผิดพลาดที่เกิดขึ้นท่ามกลางกลุ่มของการปฏิเสธสมมติฐานว่าง (R)”*

จากนิยามข้างต้น กำหนดให้ Q แทนสัดส่วนของจำนวนความผิดพลาดชนิดที่ 1 (Type I Error) หรือ False Discoveries (V) ภายใต้จำนวนการปฏิเสธสมมติฐานว่าง (H_0) หรือ Discoveries (R) ทั้งหมด นั่นคือ $Q = \frac{V}{R}$ และสามารถหาค่า FDR ได้จาก

$$FDR = Q_e = E[Q] = E\left[\frac{V}{V+S}\right] = E\left[\frac{V}{R}\right]$$

ซึ่ง $\frac{V}{R}$ จะมีค่าเท่ากับ 0 เมื่อไม่มีการปฏิเสธ สมมติฐานว่าง เมื่อ H_0 เป็นจริง

2.4.1 ขั้นตอนการควบคุม FDR ภายใต้การทดสอบที่เป็นอิสระต่อกัน (Benjamini and Hochberg 1995)

พิจารณาการทดสอบสมมติฐาน $H_1, H_2, \ldots, H_m$ โดยแต่ละการทดสอบมีค่า p-value ที่สอดคล้องคือ $P_1, P_2, \ldots, P_m$ ตามลำดับ หลังจากนั้นจึงนำค่า p-value มาเรียงลำดับจากค่าน้อยไปหาค่ามากโดยให้ $P_{(1)} \leq P_{(2)} \leq \ldots \leq P_{(m)}$ และให้ $H_{(i)}$ มีค่า p-value เท่ากับ $P_{(i)}$ เมื่อ i=1,2,...,m

ให้ k แทนค่า i ที่มีค่ามากที่สุด ซึ่งทำให้ $P_{(i)} \leq \frac{i}{m} q$ (2.15)

เราจะปฏิเสธทุกการทดสอบ $H_{(i)}$ เมื่อ i = 1, 2, ..., k

โดยการควบคุม FDR ที่ระดับ q (โดยที่ $0 \leq q \leq 1$)

ทฤษฎีบทเกี่ยวกับสถิติการทดสอบภายใต้การทดสอบค่าสัมประสิทธิ์แต่ละตัวเป็นอิสระต่อกันจากขั้นตอนข้างต้นจะควบคุม FDR ที่ระดับ q

2.4.2 ขั้นตอนการควบคุม FDR ภายใต้การทดสอบที่ไม่เป็นอิสระต่อกัน (Benjamini and Hochberg 1995)

โดย Benjamini และ Yekutieli (2001) ได้นำเสนอการพิจารณาการทดสอบ $H_1, H_2, \ldots, H_m$ โดยแต่ละการทดสอบมีค่า p-value ที่สอดคล้องคือ $P_1, P_2, \ldots, P_m$ ตามลำดับ หลังจากนั้นมีการนำค่า p-value มาเรียงลำดับจากค่าน้อยไปหาค่ามากโดยให้ $P_{(1)} \leq P_{(2)} \leq \ldots \leq P_{(m)}$ และ $H_{(i)}$ มีค่า p-value เท่ากับ $P_{(i)}$ เมื่อ i=1,2,...,m

ให้ k แทนค่า i ที่มีค่ามากที่สุด ซึ่งทำให้ $P_{(i)} \leq \frac{i}{m} q$ (2.16)

เราจะปฏิเสธทุกการทดสอบ $H_{(i)}$ เมื่อ i = 1, 2, ..., k

และจะควบคุม FDR ที่ระดับ $q \sum_{i=1}^{m} i^{-1}$

ค่า p-value ที่ได้จากวิธี Multi-Split ได้ทำการปรับค่าแล้วจากการทดสอบสมมติฐาน m สมมติฐานภายใต้ความไม่เป็นอิสระกัน ดังนั้นวิธีการคัดเลือกตัวแปรโดยใช้ FDR จะเปลี่ยนไปโดยไม่มีการหารด้วย m จะได้ว่า

$$h = \max\{i: P_{(i)} \leq iq\} \quad (2.17)$$

และเซตของตัวแปรที่ได้รับการคัดเลือกจะแสดงด้วยค่า h โดย

$$\hat{S}_{multi;FDR} = \{j: P_j \leq P_{(h)}\} \quad (2.19)$$

ถ้าไม่มีการปฏิเสธสมมติฐานว่าง แสดงว่า $\hat{S}_{multi;FDR} = \emptyset$ แต่ถ้า $P_{(i)} > iq$ จะปฏิเสธสมมติฐาน ว่างสำหรับ i=1,2,...,k ซึ่งจะควบคุม FDR ได้ที่ $q \sum_{i=1}^{m} i^{-1}$ ซึ่งเป็นระดับเดียวกับของ Benjamini และ Yekutieli

2.5 เกณฑ์ที่ใช้ในการตัดสินใจ

เกณฑ์ที่ใช้ในการตัดสินใจว่าวิธีการคัดเลือกตัวแปรวิธีการใดเหมาะสมในการคัดกรองตัวแปรสำหรับขั้นตอนวิธี Multi – Split เพื่อหาค่า p-value ในการวิเคราะห์การถดถอยของข้อมูลที่มีมิติสูงมากที่สุด โดยจะพิจารณาจากจำนวนของสัมประสิทธิ์ของตัวแปรอิสระที่มีค่าไม่เท่ากับ 0 โดยทำการประมาณขึ้นหลังจากควบคุมด้วยวิธี FDR ($|\hat{S}|$) ค่าความผิดพลาดในการตรวจจับเชิงบวก (False Positive : FP) และความผิดพลาดในการตรวจจับเชิงลบ (False Negative : FN) โดยที่

กำหนดให้ $S = \{ j : \beta_j \neq 0 \}$

$\hat{S} = \{ j :$ ปฏิเสธ $H_0 : \beta_j = 0 \}$

1. ความผิดพลาดในการตรวจจับเชิงบวก (False Positive : FP)

คือ การวัดจำนวนที่เกิดความผิดพลาดจากการปฏิเสธสมมติฐานว่างว่าค่าสัมประสิทธิ์มีค่าเท่ากับศูนย์ เมื่อค่าสัมประสิทธิ์การถดถอยที่แท้จริงมีค่าเท่ากับศูนย์สามารถคำนวณได้ดังนี้

$$FP = |\hat{S} \cap S^c| \qquad (2.18)$$

2. ความผิดพลาดในการตรวจจับเชิงลบ (False Negative : FN)

คือ การวัดจำนวนที่เกิดความผิดพลาดจากการไม่ปฏิเสธสมมติฐานว่างว่าค่าสัมประสิทธิ์มีค่าเท่ากับศูนย์ เมื่อค่าสัมประสิทธิ์การถดถอยที่แท้จริงมีค่าไม่เท่ากับศูนย์ สามารถคำนวณได้ดังนี้

$$FN = |\hat{S}^c \cap S| \qquad (2.19)$$

3. จำนวนของสัมประสิทธิ์ของตัวแปรอิสระที่มีค่าไม่เท่ากับ 0 จากการทดสอบสมมติฐาน เมื่อควบคุม FDR ($|\hat{S}|$)

บทที่ 3

วิธีการดำเนินการศึกษา

ในงานวิจัยนี้มีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพของวิธีการคัดเลือกตัวแปรและประมาณค่าสัมประสิทธิ์ระหว่างวิธี Lasso วิธี Adaptive Lasso วิธี EN และวิธี SCAD โดยมีการจำลองข้อมูลที่มีการแจกแจงแบบปกติ (Normal Distribution) หลังจากนั้นจึงทำการประมาณค่า p-value ให้กับตัวแปรอิสระที่ได้รับการเข้าเลือกเข้ามายังตัวแบบจากวิธีข้างต้น และควบคุมความผิดพลาดที่เกิดขึ้นด้วยวิธี False Discovery Rate (FDR) โดยพิจาณาประสิทธิภาพของแต่ละวิธีจากค่าความผิดพลาดในการตรวจจับเชิงบวก (False Positive : FP) ความผิดพลาดในการตรวจจับเชิงลบ (False Negative : FN) และจำนวนของสัมประสิทธิ์ของตัวแปรอิสระที่มีค่าไม่เท่ากับ 0 จากการทดสอบสมมติฐาน เมื่อควบคุม FDR โดยทำการวิเคราะห์ข้อมูลทั้งหมดโดยใช้โปรแกรม R เวอร์ชัน 3.0.3 ภายใต้ขอบเขตและวิธีการดำเนินการดังนี้

3.1 ขอบเขตของการศึกษา

ในการวิจัยครั้งนี้จะทำการศึกษาภายใต้ขอบเขตดังนี้

3.1.1 จำลองข้อมูลที่นำมาศึกษาให้มีการแจกแจงแบบปกติ (Normal Distribution) ศึกษาตัวแปรภายใต้รูปแบบความสัมพันธ์การถดถอยเชิงเส้น ในรูป

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

จาก $X \sim N(0, \Sigma)$, $\varepsilon \sim N(0, \sigma^2 I_n)$ เมื่อ $\sigma^2 = 1$

โดยที่ $\Sigma = \begin{bmatrix} \rho_{11} & \cdots & \rho_{1p} \\ \vdots & \ddots & \vdots \\ \rho_{p1} & \cdots & \rho_{pp} \end{bmatrix}$; $\rho_{ij} = \begin{cases} 1 & ; \ i = j \\ \rho^{|i-j|} & ; \ i \neq j \end{cases}$

$Y = (y_1, y_2, \ldots, y_n)'$ เป็นเวกเตอร์ของตัวแปรตามขนาด n x1

$X = (x_1, x_2, \ldots, x_n)'$ เป็นเมตริกซ์ตัวแปรอิสระขนาด n x p โดยที่ $X_i = \begin{bmatrix} x_{i1} \\ \vdots \\ x_{ip} \end{bmatrix}$

$\beta = (\beta_1, \beta_2, \ldots, \beta_p)'$ เป็นเวกเตอร์ค่าสัมประสิทธิ์การถดถอยขนาด p x 1

$\varepsilon = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)^T$ เป็นเวกเตอร์ความคลาดเคลื่อนขนาด n x 1

เมื่อ n คือขนาดตัวอย่าง

p คือจำนวนตัวแปรอิสระ

3.1.2 อัตราส่วนของขนาดตัวอย่างต่อจำนวนตัวแปรอิสระ (n : p) ที่ 1:2, 1:5 และ 1:10

กรณีที่ 1 : n = 10 เปรียบเทียบอัตราส่วน n:p ที่ 10 : 20, 10 : 50, 10 : 100

กรณีที่ 2 : n = 100 เปรียบเทียบอัตราส่วน n:p ที่ 100 : 200, 100 : 500, 100 : 1,000

กรณีที่ 3 : n = 200 เปรียบเทียบอัตราส่วน n:p ที่ 200 : 400, 200 : 1,000, 200 : 2,000

3.1.3 ลักษณะของขนาด (Effect Size) ของค่าสัมประสิทธิ์ตัวที่ไม่เท่ากับ 0 โดยแบ่งเป็นขนาดเล็ก $(0 < |\beta| < 1)$ และขนาดใหญ่ $(1 < |\beta| < 10)$

3.1.4 ร้อยละของจำนวนสัมประสิทธิ์ (β) ที่ไม่เท่ากับ 0 เมื่อเทียบกับขนาดตัวอย่างที่ร้อยละ 10, 20 และ 50

3.1.5 ศึกษาภายใต้ความสัมพันธ์ (Correlation) ของตัวแปรอิสระ 3 ระดับ โดยมีจำนวนตัวแปรอิสระ(p) ที่ใช้ในงานวิจัยทั้งหมด 8 ระดับ คือ 20, 50, 100, 200, 400, 500, 1000, 2000 ตัว จะมีเมตริกซ์ความแปรปรวนร่วมของแต่ละระดับความสัมพันธ์ ดังนี้

เมตริกซ์ความแปรปรวนร่วม $(\boldsymbol{\Sigma}) = \begin{bmatrix} \rho_{11} & \rho_{12} & \cdots & \rho_{1p} \\ \rho_{21} & \rho_{22} & \cdots & \rho_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{p1} & \rho_{p2} & \cdots & \rho_{pp} \end{bmatrix}_{pxp}$

โดยที่ $\rho_{ij} = \begin{cases} 1 & ; \ i = j \\ \rho^{|i-j|} & ; \ i \neq j \end{cases}$

โดยความสัมพันธ์ (Correlation) ของตัวแปรอิสระทั้ง 3 ระดับ คือ

ระดับที่ 1 : $\rho = 0$

ระดับที่ 2 : $\rho = 0.5$

ระดับที่ 3 : $\rho = 0.9$

<u>กรณีที่มี $\rho = 0$</u>

ที่จำนวนตัวแปรอิสระทั้งหมด 8 ระดับ คือ 20, 50, 100, 200, 400, 500, 1000, 2000 ตัว

จะได้ว่า $\rho_{ij} = 0$ เท่ากันในทุกกรณี

<u>กรณีที่มี $\rho = 0.5$</u>

ที่จำนวนตัวแปรอิสระเท่ากับ 20 ค่า จะได้ว่า $\rho_{ij} \in [1.9\times10^{-6},0.5]$; $i \neq j$

ที่จำนวนตัวแปรอิสระเท่ากับ 50 ค่า จะได้ว่า $\rho_{ij} \in [1.7\times10^{-15},0.5]$; $i \neq j$

ที่จำนวนตัวแปรอิสระเท่ากับ 100 ค่า จะได้ว่า $\rho_{ij} \in [1.6\times10^{-30},0.5]$; $i \neq j$

ที่จำนวนตัวแปรอิสระเท่ากับ 200 ค่า จะได้ว่า $\rho_{ij} \in [1.2\times10^{-60},0.5]$; $i \neq j$

ที่จำนวนตัวแปรอิสระเท่ากับ 400 ค่า จะได้ว่า $\rho_{ij} \in [0,0.5]$; $i \neq j$

ที่จำนวนตัวแปรอิสระเท่ากับ 500 ค่า จะได้ว่า $\rho_{ij} \in [0,0.5]$; $i \neq j$

ที่จำนวนตัวแปรอิสระเท่ากับ 1000 ค่า จะได้ว่า $\rho_{ij} \in [0,0.5]$; $i \neq j$

ที่จำนวนตัวแปรอิสระเท่ากับ 2000 ค่า จะได้ว่า $\rho_{ij} \in [0,0.5]$; $i \neq j$

<u>กรณีที่มี $\rho = 0.9$</u>

ที่จำนวนตัวแปรอิสระเท่ากับ 20 ค่า จะได้ว่า $\rho_{ij} \in [0.14,0.9]$; $i \neq j$

ที่จำนวนตัวแปรอิสระเท่ากับ 50 ค่า จะได้ว่า $\rho_{ij} \in [0.0057,0.9]$; $i \neq j$

ที่จำนวนตัวแปรอิสระเท่ากับ 100 ค่า จะได้ว่า $\rho_{ij} \in [0.000029,0.9]$; $i \neq j$

ที่จำนวนตัวแปรอิสระเท่ากับ 200 ค่า จะได้ว่า $\rho_{ij} \in [7.8\times10^{-10},0.9]$; $i \neq j$

ที่จำนวนตัวแปรอิสระเท่ากับ 400 ค่า จะได้ว่า $\rho_{ij} \in [5.5\times10^{19},0.9]$; $i \neq j$

ที่จำนวนตัวแปรอิสระเท่ากับ 500 ค่า จะได้ว่า $\rho_{ij} \in [1.4\times10^{-23},0.9]$; $i \neq j$

ที่จำนวนตัวแปรอิสระเท่ากับ 1000 ค่า จะได้ว่า $\rho_{ij} \in [1.9\times10^{-46},0.9]$; $i \neq j$

ที่จำนวนตัวแปรอิสระเท่ากับ 2000 ค่า จะได้ว่า $\rho_{ij} \in [3.4\times10^{-92},0.9]$; $i \neq j$

ตารางที่ 3.1.1 แสดงค่าจำนวนสัมประสิทธิ์ (β) ที่ไม่เท่ากับ 0 ตามขอบเขตของข้อมูลในกรณีที่ขนาดตัวอย่าง(n) เท่ากับ 10 และมีขนาด (Effect size) ของค่าสัมประสิทธิ์ตัวที่ไม่เท่ากับ 0 ที่ขนาดเล็ก $(0 < |\beta| < 1)$ และขนาดใหญ่ $(1 < |\beta| < 10)$ ของวิธี Lasso, Adaptive Lasso, EN และ SCAD

ρ	n : p	ร้อยละของจำนวนสัมประสิทธิ์ที่ไม่เท่ากับ 0 เมื่อเทียบกับขนาดตัวอย่าง	จำนวนสัมประสิทธิ์ที่ไม่เท่ากับ 0	
			$0 < \|\beta\| < 1$	$1 < \|\beta\| < 10$
0	10 : 20	10	1	1
		20	2	2
		50	5	5
	10 : 50	10	1	1
		20	2	2
		50	5	5
	10 : 100	10	1	1
		20	2	2
		50	5	5
0.5	10 : 20	10	1	1
		20	2	2
		50	5	5
	10 : 50	10	1	1
		20	2	2
		50	5	5
	10 : 100	10	1	1
		20	2	2
		50	5	5
0.9	10 : 20	10	1	1
		20	2	2
		50	5	5
	10 : 50	10	1	1
		20	2	2
		50	5	5
	10 : 100	10	1	1
		20	2	2
		50	5	5

ตารางที่ 3.1.2 แสดงค่าจำนวนสัมประสิทธิ์ (β) ที่ไม่เท่ากับ 0 ตามขอบเขตของข้อมูลในกรณีที่ขนาดตัวอย่าง(n) เท่ากับ 100 และมีขนาด (Effect size) ของค่าสัมประสิทธิ์ตัวที่ไม่เท่ากับ 0 ที่ขนาดเล็ก $(0 < |\beta| < 1)$ และขนาดใหญ่ $(1 < |\beta| < 10)$ ของวิธี Lasso, Adaptive Lasso, EN และ SCAD

ρ	n : p	ร้อยละของจำนวนสัมประสิทธิ์ที่ไม่เท่ากับ 0 เมื่อเทียบกับขนาดตัวอย่าง	จำนวนสัมประสิทธิ์ที่ไม่เท่ากับ 0	
			$0 < \|\beta\| < 1$	$1 < \|\beta\| < 10$
0	100 : 200	10	10	10
		20	20	20
		50	50	50
	100 : 500	10	10	10
		20	20	20
		50	50	50
	100 : 1000	10	10	10
		20	20	20
		50	50	50
0.5	100 : 200	10	10	10
		20	20	20
		50	50	50
	100 : 500	10	10	10
		20	20	20
		50	50	50
	100 : 1000	10	10	10
		20	20	20
		50	50	50
0.9	100 : 200	10	10	10
		20	20	20
		50	50	50
	100 : 500	10	10	10
		20	20	20
		50	50	50
	100 : 1000	10	10	10
		20	20	20
		50	50	50

ตารางที่ 3.1.3 แสดงค่าจำนวนสัมประสิทธิ์ (β) ที่ไม่เท่ากับ 0 ตามขอบเขตของข้อมูลในกรณีที่ขนาดตัวอย่าง(n) เท่ากับ 200 และมีขนาด (Effect size) ของค่าสัมประสิทธิ์ตัวที่ไม่เท่ากับ 0 ที่ขนาดเล็ก $(0 < |\beta| < 1)$ และขนาดใหญ่ $(1 < |\beta| < 10)$ ของวิธี Lasso, Adaptive Lasso, EN และ SCAD

ρ	n : p	ร้อยละของจำนวนสัมประสิทธิ์ที่ไม่เท่ากับ 0 เมื่อเทียบกับขนาดตัวอย่าง	จำนวนสัมประสิทธิ์ที่ไม่เท่ากับ 0	
			$0 < \|\beta\| < 1$	$1 < \|\beta\| < 10$
0	200 : 400	10	20	20
		20	40	40
		50	100	100
	200 : 1000	10	20	20
		20	40	40
		50	100	100
	200 : 2000	10	20	20
		20	40	40
		50	100	100
0.5	200 : 400	10	20	20
		20	40	40
		50	100	100
	200 : 1000	10	20	20
		20	40	40
		50	100	100
	200 : 2000	10	20	20
		20	40	40
		50	100	100
0.9	200 : 400	10	20	20
		20	40	40
		50	100	100
	200 : 1000	10	20	20
		20	40	40
		50	100	100
	200 : 2000	10	20	20
		20	40	40
		50	100	100

3.2 ขั้นตอนในการดำเนินการศึกษา

1. ศึกษาตัวแบบและทฤษฎีที่เกี่ยวข้อง
2. กำหนดและจำลองข้อมูล
   2.1 กำหนดค่าเริ่มต้นโดยการสร้างข้อมูลที่มีจำนวนค่าสังเกต n ค่า และจำนวนพารามิเตอร์ p ตัว โดยใช้อัตราส่วน n:p คือ
      - 10:20, 10:50 และ 10:100
      - 100:200, 100:500 และ 100:1000
      - 200:400, 200:1000 และ 200:2000

   2.2 กำหนดให้ร้อยละของจำนวนสัมประสิทธิ์ (β) ที่ไม่เท่ากับ 0 เมื่อเทียบกับขนาดตัวอย่างที่ 10, 20 และ 50

   2.3 กำหนดให้ขนาด (Effect size) ของสัมประสิทธิ์ที่ไม่เท่ากับ 0 ในกรณี ขนาดเล็ก $(0 < |\beta| < 1)$ และขนาดใหญ่ $(1 < |\beta| < 10)$

   2.4 จำลองข้อมูลภายใต้รูปแบบความสัมพันธ์การถดถอยเชิงเส้น ในรูป

$$Y = \mathbf{X}\beta + \varepsilon$$

จาก $\mathbf{X} \sim N(0, \mathbf{\Sigma})$, $\varepsilon \sim N(0, \sigma^2 \boldsymbol{I}_n)$ เมื่อ $\sigma^2 = 1$

โดยที่ $\mathbf{\Sigma} = \begin{bmatrix} \rho_{11} & \cdots & \rho_{1p} \\ \vdots & \ddots & \vdots \\ \rho_{p1} & \cdots & \rho_{pp} \end{bmatrix}$; $\rho_{ij} = \begin{cases} 1 & ; \ i = j \\ \rho^{|i-j|} & ; \ i \neq j \end{cases}$

$\mathbf{Y} = (y_1, y_2, \ldots, y_n)'$ เป็นเวกเตอร์ของตัวแปรตามขนาด n x1

$\mathbf{X} = (x_1, x_2, \ldots, x_n)'$ เป็นเมตริกซ์ตัวแปรอิสระขนาด n x p โดยที่ $X_i = \begin{bmatrix} x_{i1} \\ \vdots \\ x_{ip} \end{bmatrix}$

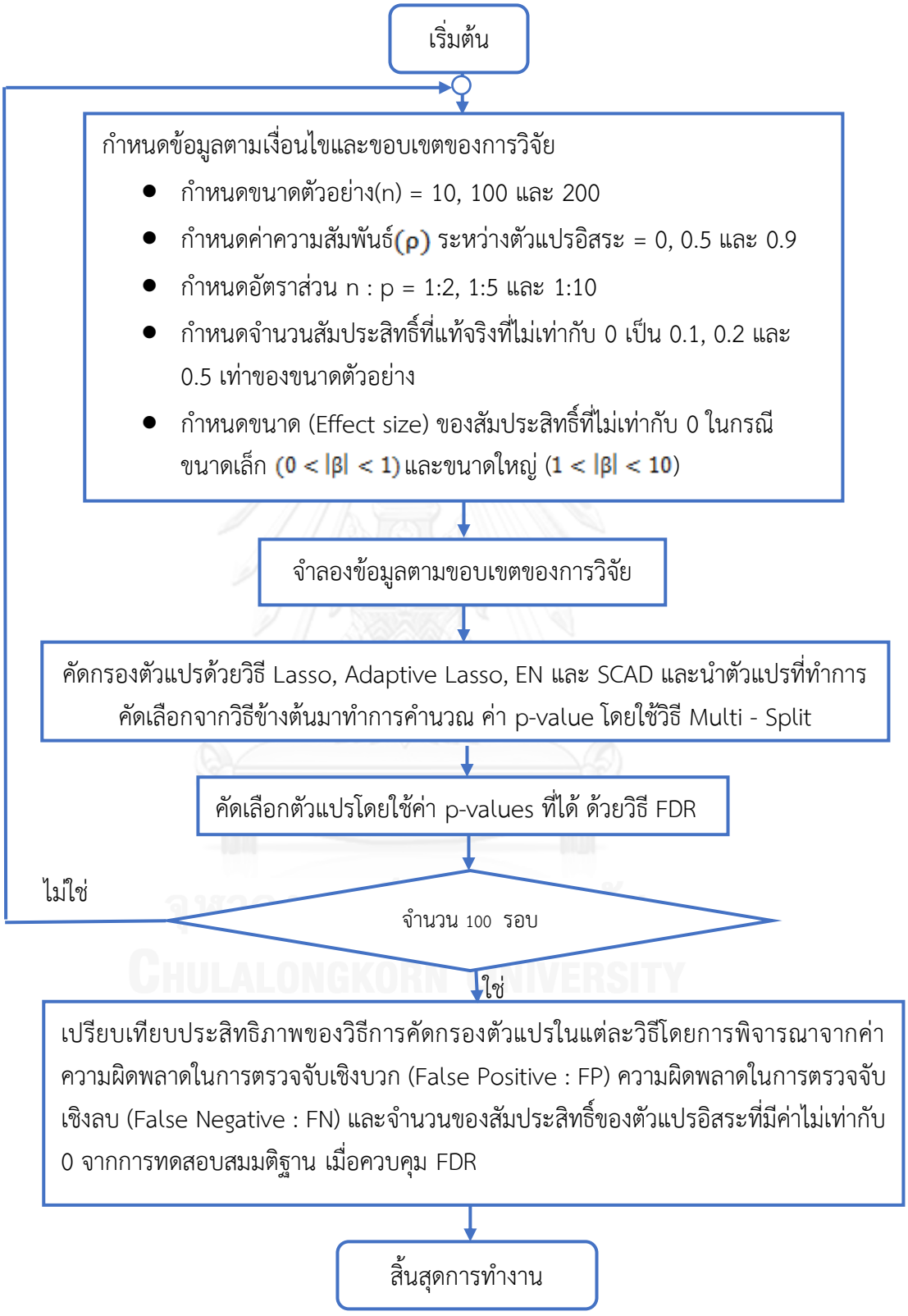
$\beta = (\beta_1, \beta_2, \ldots, \beta_p)'$ เป็นเวกเตอร์ค่าสัมประสิทธิ์การถดถอยขนาด p x 1

$\mathcal{E} = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)'$ เป็นเวกเตอร์ความคลาดเคลื่อนขนาด n x 1

เมื่อ n คือขนาดตัวอย่าง
p คือจำนวนตัวแปรอิสระ

3. นำข้อมูลที่จำลองขึ้นมาศึกษาและใช้วิธี Lasso, Adaptive Lasso, EN, SCAD ในขั้นตอนการคัดเลือกข้อมูลสำหรับวิธีการ Multi-Sample Split ในการคำนวณค่า p-value
4. นำค่า p-value ที่ได้จากการประมาณค่าวิธีต่างๆในข้อ 3. มาคัดกรองตัวแปรในขั้นสุดท้ายด้วยวิธีการควบคุม False Discovery Rate (FDR)
5. นำข้อมูลที่ได้จากข้อ 4 มาหาค่า FP, FN และ เซต $\hat{S}$
6. วิเคราะห์ผลลัพธ์โดยทำการเปรียบเทียบค่า FP FN และ $|\hat{S}|$ ที่ได้จากการคัดเลือกตัวแปรในแต่ละวิธีเมื่อขนาดตัวอย่าง (n) เท่ากับ 10, 100 และ 200 และขนาดของ Effect size เป็นขนาดเล็ก ($0 < |\beta| < 1$) และ ขนาดใหญ่ ($1 < |\beta| < 10$) โดยจำแนกตามอัตราส่วนขนาดตัวอย่างต่อจำนวนตัวแปร (n:p), ขนาด(Effect size) ของสัมประสิทธิ์ที่ไม่เท่ากับ 0, จำนวนสัมประสิทธิ์ (β) ที่ไม่เท่ากับ 0 และความสัมพันธ์ (Correlation) ของตัวแปรอิสระ

3.3 ขั้นตอนการทำงานของโปรแกรม



บทที่ 4

ผลการวิจัย

งานวิจัยนี้มีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพของวิธีการคัดเลือกตัวแปรและประมาณค่าสัมประสิทธิ์ระหว่างวิธี Lasso วิธี Adaptive Lasso วิธี EN และวิธี SCAD โดยจะพิจารณาแยกตามขนาดตัวอย่างที่ n = 10, 100 และ 200 และอัตราส่วนระหว่างขนาดของตัวอย่างและจำนวนตัวแปรอิสระ โดยมีเกณฑ์ที่ใช้ในการพิจาณาประสิทธิภาพของแต่ละวิธีจากค่าความผิดพลาดในการตรวจจับเชิงบวก (False Positive : FP) ความผิดพลาดในการตรวจจับเชิงลบ (False Negative : FN) และจำนวนของสัมประสิทธิ์ของตัวแปรอิสระที่มีค่าไม่เท่ากับ 0 จากการทดสอบสมมติฐานโดยเฉลี่ย เมื่อควบคุม FDR ($|\hat{S}|$) โดยถ้าวิธีใดให้ค่าวัดประสิทธิภาพ FP และ FN ต่ำที่สุดและมีค่าเข้าใกล้ 0 และมีจำนวนสัมประสิทธิ์ที่ไม่เท่ากับ 0 จากการทดสอบสมมติฐานมีขนาดใกล้เคียงกับตัวแบบที่แท้จริงมากที่สุดจะถือได้ว่าเป็นวิธีที่มีประสิทธิภาพและมีความเหมาะสมในการคัดเลือกตัวแปรสำหรับข้อมูลที่มีมิติสูง(High-Dimensional Data) มากที่สุด

อักษรย่อและสัญลักษณ์ต่างๆที่ปรากฏในการนำเสนอผลการวิจัยทั้งในตารางและข้อความต่างๆแทนความหมายดังนี้

n	แทน ขนาดของตัวอย่าง
p	แทน จำนวนตัวแปรอิสระ
ρ	แทน ความสัมพันธ์ (Correlation) ของตัวแปรอิสระ
n: p	แทน ขนาดของตัวอย่างต่อจำนวนตัวแปรอิสระ
Effect Size	แทน ขนาดของสัมประสิทธิ์ที่ไม่เท่ากับ 0
Small Size	แทน ขนาดของสัมประสิทธิ์ที่ไม่เท่ากับกับ 0 ที่มีค่าอยู่ระหว่าง -1 ถึง 1 $(0 < \|\beta\| < 1)$
Large Size	แทน ขนาดของสัมประสิทธิ์ที่ไม่เท่ากับ 0 ที่มีค่ามากกว่า -10 และน้อยกว่า -1 หรือมากกว่า 1 และน้อยกว่า 10 $(1 < \|\beta\| < 10)$
Lasso	แทน การคัดเลือกตัวแปรด้วยวิธี Lasso
Adaptive Lasso	แทน การคัดเลือกตัวแปรด้วยวิธี Adaptive Lasso
EN	แทน การคัดเลือกตัวแปรด้วยวิธี EN
SCAD	แทน การคัดเลือกตัวแปรด้วยวิธี SCAD
FP	แทน การตรวจจับเชิงบวก (False Positive)

FN	แทน การตรวจจับเชิงลบ (False Negative)
$\|S\|$	แทน จำนวนของสัมประสิทธิ์ของตัวแปรอิสระที่มีค่าไม่เท่ากับ 0 ของตัวแบบที่แท้จริง
$\|\hat{S}\|$	แทน จำนวนของสัมประสิทธิ์ของตัวแปรอิสระที่มีค่าไม่เท่ากับ 0 จากการทดสอบสมมติฐานเมื่อควบคุม FDR ระดับ 0.1
β	แทน สัมประสิทธิ์การถดถอยของตัวแปรอิสระ
Mean	แทน ค่าเฉลี่ย
S.D.	แทน ค่าเบี่ยงเบนมาตรฐาน

สำหรับงานวิจัยนี้จะนำเสนอผลการเปรียบเทียบโดยแบ่งออกเป็น 3 ส่วน คือ ในส่วนที่ 1 จะเปรียบเทียบความผิดพลาดในการตรวจจับเชิงบวก (FP) ความผิดพลาดในการตรวจจับเชิงลบ (FN) และจำนวนของสัมประสิทธิ์ของตัวแปรอิสระที่มีค่าไม่เท่ากับ 0 จากการทดสอบสมมติฐาน เมื่อควบคุม FDR ที่ระดับ 0.1 ของข้อมูลจำลองขนาด 10 ระหว่างการคัดกรองตัวแปรจากวิธี Lasso, Adaptive Lasso, EN และ SCAD ส่วนที่ 2 จะเปรียบเทียบความผิดพลาดในการตรวจจับเชิงบวก (FP) ความผิดพลาดในการตรวจจับเชิงลบ (FN) และจำนวนของสัมประสิทธิ์ของตัวแปรอิสระที่มีค่าไม่เท่ากับ 0 จากการทดสอบสมมติฐาน เมื่อควบคุม FDR ที่ระดับ 0.1 ของข้อมูลจำลองขนาด 100 ระหว่างการคัดกรองตัวแปรจากทั้ง 4 วิธีข้างต้นและส่วนที่ 3 จะเปรียบเทียบความผิดพลาดในการตรวจจับเชิงบวก (FP) ความผิดพลาดในการตรวจจับเชิงลบ (FN) และจำนวนของสัมประสิทธิ์ของตัวแปรอิสระที่มีค่าไม่เท่ากับ 0 จากการทดสอบสมมติฐาน เมื่อควบคุม FDR ที่ระดับ 0.1 ของข้อมูลจำลองขนาด 200 จากการทดสอบสมมติฐานโดยเฉลี่ย เมื่อควบคุม FDR ระหว่างการคัดกรองตัวแปรจากทั้ง 4 วิธีข้างต้น

โดยผลการวิจัยจะแบ่งออกเป็น 3 ส่วน ดังนี้

<u>ส่วนที่ 1</u> ผลการเปรียบเทียบความผิดพลาดในการตรวจจับเชิงบวก (FP) ความผิดพลาดในการตรวจจับเชิงลบ (FN) และจำนวนของสัมประสิทธิ์ของตัวแปรอิสระที่มีค่าไม่เท่ากับ 0 จากการทดสอบสมมติฐาน เมื่อควบคุม FDR ที่ระดับ 0.1 ของข้อมูลจำลองขนาด 10 ระหว่างการคัดกรองตัวแปรจากวิธี Lasso, Adaptive Lasso, EN และ SCAD เมื่อพิจารณาในกรณี

1.1 เมื่อกำหนดให้อัตราส่วนของขนาดตัวอย่างต่อจำนวนตัวแปรอิสระ (n : p) ที่ 10:20, 10:50, 10:100

1.2 เมื่อกำหนดให้ขนาด (Effect Size) ของค่าสัมประสิทธิ์ตัวที่ไม่เท่ากับ 0 โดยแบ่งเป็นขนาดเล็ก $(0 < |\beta| < 1)$ และขนาดใหญ่ $(1 < |\beta| < 10)$

1.3 เมื่อกำหนดให้ร้อยละของจำนวนสัมประสิทธิ์ (β) ที่ไม่เท่ากับ 0 เมื่อเทียบกับขนาดตัวอย่างที่ร้อยละ 10, 20 และ 50

1.4 เมื่อกำหนดให้ความสัมพันธ์ (Correlation) ของตัวแปรอิสระ 3 ระดับ คือ $\rho = 0$ $\rho = 0.5$ และ $\rho = 0.9$

<u>ส่วนที่ 2</u> ผลเปรียบเทียบความผิดพลาดในการตรวจจับเชิงบวก (FP) ความผิดพลาดในการตรวจจับเชิงลบ (FN) และจำนวนของสัมประสิทธิ์ของตัวแปรอิสระที่มีค่าไม่เท่ากับ 0 จากการทดสอบสมมติฐาน เมื่อควบคุม FDR ที่ระดับ 0.1 ของข้อมูลจำลองขนาด 100 ระหว่างการคัดกรองตัวแปรจากวิธี Lasso, Adaptive Lasso, EN และ SCAD เมื่อพิจารณาในกรณี

1.1 เมื่อกำหนดให้อัตราส่วนของขนาดตัวอย่างต่อจำนวนตัวแปรอิสระ (n : p) ที่ 100:200, 100:500, 100:1000

1.2 เมื่อกำหนดให้ขนาด (Effect Size) ของค่าสัมประสิทธิ์ตัวที่ไม่เท่ากับ 0 โดยแบ่งเป็นขนาด

เล็ก $(0 < |\beta| < 1)$ และขนาดใหญ่ $(1 < |\beta| < 10)$

1.3 เมื่อกำหนดให้ร้อยละของจำนวนสัมประสิทธิ์ (β) ที่ไม่เท่ากับ 0 เมื่อเทียบกับขนาดตัวอย่างที่ร้อยละ 10, 20 และ 50

1.4 เมื่อกำหนดให้ความสัมพันธ์ (Correlation) ของตัวแปรอิสระ 3 ระดับ คือ $\rho = 0$ $\rho = 0.5$ และ $\rho = 0.9$

<u>ส่วนที่ 3</u> ผลเปรียบเทียบความผิดพลาดในการตรวจจับเชิงบวก (FP) ความผิดพลาดในการตรวจจับเชิงลบ (FN) และจำนวนของสัมประสิทธิ์ของตัวแปรอิสระที่มีค่าไม่เท่ากับ 0 จากการทดสอบสมมติฐาน เมื่อควบคุม FDR ที่ระดับ 0.1 ของข้อมูลจำลองขนาด 200 ระหว่างการคัดกรองตัวแปรจากวิธี Lasso, Adaptive Lasso, EN และ SCAD เมื่อพิจารณาในกรณี

1.1 เมื่อกำหนดให้อัตราส่วนของขนาดตัวอย่างต่อจำนวนตัวแปรอิสระ (n : p) ที่ 200:400, 200:1000, 200:2000

1.2 เมื่อกำหนดให้ขนาด (Effect Size) ของค่าสัมประสิทธิ์ตัวที่ไม่เท่ากับ 0 โดยแบ่งเป็นขนาดเล็ก $(0 < |\beta| < 1)$ และขนาดใหญ่ $(1 < |\beta| < 10)$

1.3 เมื่อกำหนดให้ร้อยละของจำนวนสัมประสิทธิ์ (β) ที่ไม่เท่ากับ 0 เมื่อเทียบกับขนาดตัวอย่างที่ร้อยละ 10, 20 และ 50

1.4 เมื่อกำหนดให้ความสัมพันธ์ (Correlation) ของตัวแปรอิสระ 3 ระดับ คือ $\rho = 0$ $\rho = 0.5$ และ $\rho = 0.9$

4.1 ผลการเปรียบเทียบค่าความผิดพลาดในการตรวจจับเชิงบวก (FP) ค่าความผิดพลาดในการตรวจจับเชิงลบ (FN) และจำนวนของสัมประสิทธิ์ของตัวแปรอิสระที่มีค่าไม่เท่ากับ 0 จากการทดสอบสมมติฐานโดยเฉลี่ย เมื่อควบคุม FDR ที่ระดับ 0.1 ของข้อมูลจำลองขนาด 10 ระหว่างการคัดกรองตัวแปรด้วยวิธี Lasso, Adaptive Lasso, EN และ SCAD

ในส่วนนี้ผู้วิจัยต้องการศึกษาเปรียบเทียบการคัดกรองตัวแปรจากวิธี Lasso, Adaptive Lasso, EN และ SCAD และเพื่อพิจารณาว่าปัจจัยใดที่ส่งผลต่อประสิทธิภาพการทำงานของวิธีการคัดกรองตัวแปรแต่ละวิธี ภายใต้ปัจจัย ดังต่อไปนี้

1. อัตราส่วนระหว่างขนาดตัวอย่างและจำนวนตัวแปรอิสระ คือ 1:2, 1:5 และ 1:10
2. ร้อยละของจำนวนสัมประสิทธิ์ที่ไม่เท่ากับ 0 $(\beta \neq 0)$ เมื่อเทียบกับขนาดตัวอย่าง คือ ร้อยละ 10, 20 และ 50
3. ขนาด (Effect Size) ของสัมประสิทธิ์ที่ไม่เท่ากับ 0 คือ Small Size $(0 < |\beta| < 1)$ และ Large Size $(1 < |\beta| < 10)$
5. ความสัมพันธ์(Correlation) ของตัวแปรอิสระ คือ ρ = 0, 0.5, 0.9

โดยแสดงผลในตารางที่ 4.1.1 - 4.1.6 โดยแต่ละตารางมีรายละเอียดดังนี้

เกณฑ์ที่ใช้ในการวัด	ตารางที่	ปัจจัยที่ใช้ในการพิจารณา	วิธีการคัดกรองตัวแปรที่ต้องการเปรียบเทียบ
$\|\hat{S}\|$	4.1.1	• n ; p • จำนวนสัมประสิทธิ์ของโมเดลที่แท้จริงที่มีค่าไม่เท่ากับ 0 ($\beta \neq 0$) • ความสัมพันธ์ (Correlation) ของตัวแปรอิสระ • ขนาด(Effect Size) ของสัมประสิทธิ์ที่ไม่เท่ากับ 0	1. Lasso 2. Adaptive Lasso 3. EN 4. SCAD
	4.1.2		
FP	4.1.3		
	4.1.4		
FN	4.1.5		
	4.1.6		

ตารางที่ 4.1.1 แสดงค่าเฉลี่ย(ค่าเบี่ยงเบนมาตรฐาน) ของ |Ŝ| เมื่อควบคุม FDR ที่ระดับ 0.1 โดยคำนวณจากข้อมูล 100 ชุด กรณีที่ขนาดตัวอย่าง (n) เท่ากับ 10 และขนาดของสัมประสิทธิ์ที่ไม่เท่ากับ 0 มีขนาดเล็ก (Small effect Size)

n : p	|S|	ขนาด(Effect size) ของสัมประสิทธิ์ที่ไม่เท่ากับ 0 มีขนาดเล็ก $(0 < \|\beta\| < 1)$								
		ความสัมพันธ์ (Correlation) ของตัวแปรอิสระ								
		$\rho = 0$			$\rho = 0.5$			$\rho = 0.9$		
		Lasso	Adaptive Lasso	SCAD	Lasso	Adaptive Lasso	SCAD	Lasso	Adaptive Lasso	SCAD
10:20	1	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)
	2	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0.010 (0.100)	0.010 (0.100)
	5	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0.010 (0.100)	0 (0.000)
10:50	1	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)
	2	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)
	5	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)
10:100	1	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)
	2	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)
	5	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)

หมายเหตุ ช่องที่ระบายสี หมายถึง วิธีที่เหมาะสมที่สุดในแต่ละกรณี

จากตารางที่ 4.1.1 ซึ่งแสดงผลของ $|\hat{S}|$ โดยเฉลี่ยของข้อมูลจำลองขนาด 10 ระหว่างการคัดกรองตัวแปรด้วยวิธี Lasso, Adaptive Lasso, En และ SCAD โดยเปรียบเทียบกับจำนวนสัมประสิทธิ์ของตัวแบบที่แท้จริง ที่ขนาดของสัมประสิทธิ์ไม่เท่ากับ 0 เป็นขนาดเล็ก $(0 < |\beta| < 1)$ พบว่า

1. ที่จำนวนตัวแปรอิสระ (p) เท่ากับ 20
   - เมื่อ $|S|$ มี 1 ค่า ทุกระดับความสัมพันธ์(Correlation) ของตัวแปรอิสระ การคัดกรองตัวแปรด้วย 4 วิธีข้างต้นหาค่า $|\hat{S}|$ ได้เท่ากับ 0 เท่ากันทุกวิธี นั่นแสดงว่าไม่มีวิธีคัดกรองตัวแปรที่ให้ค่าสัมประสิทธิ์ตัวใดไม่เท่ากับ 0 จากการทดสอบสมมติฐาน
   - เมื่อ $|S|$ มี 2 ค่า เมื่อตัวแปรอิสระมีค่าความสัมพันธ์ที่ $\rho = 0$ และ 0.5 การคัดกรองตัวแปรด้วย 4 วิธีข้างต้นหาค่า $|\hat{S}|$ ได้เท่ากับ 0 เท่ากันทุกวิธี นั่นแสดงว่าไม่มีวิธีคัดกรองตัวแปรที่ให้ค่าสัมประสิทธิ์ตัวใดไม่เท่ากับ 0 จากการทดสอบสมมติฐาน แต่เมื่อตัวแปรอิสระมีความสัมพันธ์ที่ $\rho = 0.9$ การคัดกรองตัวแปรด้วยวิธี Adaptive Lasso และวิธี SCAD มีค่า $|\hat{S}|$ ใกล้เคียงกับ $|S|$ มากที่สุด ดังนั้นการคัดกรองตัวแปรด้วยวิธี Adaptive Lasso และวิธี SCAD จึงเป็นวิธีที่เหมาะสมที่สุด
   - เมื่อ $|S|$ มี 5 ค่า เมื่อตัวแปรอิสระมีค่าความสัมพันธ์ที่ $\rho = 0$ และ 0.5 การคัดกรองตัวแปรด้วย 4 วิธีข้างต้นหาค่า $|\hat{S}|$ ได้เท่ากับ 0 เท่ากันทุกวิธี นั่นแสดงว่าไม่มีวิธีคัดกรองตัวแปรที่ให้ค่าสัมประสิทธิ์ตัวใดไม่เท่ากับ 0 จากการทดสอบสมมติฐาน แต่เมื่อตัวแปรอิสระมีความสัมพันธ์กันที่ $\rho = 0.9$ การคัดกรองตัวแปรด้วยวิธี Adaptive Lasso มีค่า $|\hat{S}|$ ใกล้เคียงกับ $|S|$ มากที่สุด ที่สุด ดังนั้นการคัดกรองตัวแปรด้วยวิธี Adaptive Lasso จึงเป็นวิธีที่เหมาะสมที่สุด
2. ที่จำนวนตัวแปรอิสระ (p) เท่ากับ 50 และ 100
   - เมื่อ $|S|$ มี 1, 2 และ 5 ค่า ตัวแปรอิสระมีค่าความสัมพันธ์ที่ $\rho = 0, 0.5$ และ 0.9 การประมาณค่าด้วย 4 วิธีข้างต้นหาค่า $|\hat{S}|$ ได้เท่ากับ 0 เท่ากันทุกวิธี นั่นแสดงว่าไม่มีวิธีคัดกรองตัวแปรที่ให้ค่าสัมประสิทธิ์ตัวใดไม่เท่ากับ 0 จากการทดสอบสมมติฐาน

และจากผลในตารางที่ 4.1.1 ยังสามารถสรุปได้ว่า

- จำนวนของ p ไม่มีผลต่อประสิทธิภาพในการหาค่า $|\hat{S}|$ ให้ใกล้เคียงกับ $|S|$
- ขนาดของ ρ ไม่มีผลต่อประสิทธิภาพในการหาค่า $|\hat{S}|$ ให้ใกล้เคียงกับ $|S|$
- ขนาดของ $|S|$ มีไม่มีผลต่อประสิทธิภาพในการหาค่า $|\hat{S}|$ ให้ใกล้เคียงกับ $|S|$

ตารางที่ 4.1.2 แสดงค่าเฉลี่ย(ค่าเบี่ยงเบนมาตรฐาน) ของ $|\hat{S}|$ เมื่อควบคุม FDR ที่ระดับ 0.1 โดยคำนวณจากข้อมูล 100 ชุด กรณีที่ขนาดตัวอย่าง (n) เท่ากับ 10 และขนาดของสัมประสิทธิ์ที่ไม่เท่ากับ 0 มีขนาดใหญ่ (Large effect Size)

n : p	$\|S\|$	ขนาด(Effect size) ของสัมประสิทธิ์ที่ไม่เท่ากับ 0 มีขนาดใหญ่ $(1 < \|\beta\| < 10)$								
		ความสัมพันธ์ (Correlation) ของตัวแปรอิสระ								
		$\rho = 0$			$\rho = 0.5$			$\rho = 0.9$		
		Lasso	Adaptive Lasso	SCAD	Lasso	Adaptive Lasso	SCAD	Lasso	Adaptive Lasso	SCAD
10:20	1	0.450 (0.500)	0.480 (0.502)	0.420 (0.496)	0.530 (0.502)	0.580 (0.496)	0.460 (0.501)	0.410 (0.494)	0.590 (0.494)	0.540 (0.501)
	2	0.080 (0.273)	0.100 (0.302)	0.070 (0.256)	0.030 (0.171)	0.070 (0.293)	0.050 (0.219)	0.140 (0.349)	0.220 (0.416)	0.140 (0.349)
	5	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0.010 (0.100)
10:50	1	0.560 (0.500)	0.570 (0.498)	0.460 (0.501)	0.400 (0.492)	0.420 (0.496)	0.390 (0.490)	0.320 (0.469)	0.390 (0.490)	0.340 (0.476)
	2	0.020 (0.141)	0.070 (0.256)	0.030 (0.171)	0 (0.000)	0.020 (0.141)	0 (0.000)	0.070 (0.256)	0.090 (0.288)	0.070 (0.256)
	5	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)
10:100	1	0.300 (0.461)	0.340 (0.476)	0.250 (0.435)	0.360 (0.482)	0.420 (0.496)	0.340 (0.476)	0.330 (0.473)	0.350 (0.479)	0.290 (0.456)
	2	0.020 (0.141)	0.020 (0.141)	0.010 (0.100)	0.010 (0.100)	0.020 (0.141)	0.020 (0.141)	0.040 (0.197)	0.050 (0.219)	0.020 (0.141)
	5	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0.010 (0.100)	0.010 (0.100)

หมายเหตุ ช่องที่ระบายสี หมายถึง วิธีที่เหมาะสมที่สุดในแต่ละกรณี

จากตารางที่ 4.1.2 ซึ่งแสดงผลของ $|\hat{S}|$ โดยเฉลี่ยของข้อมูลจำลองขนาด 10 ระหว่างการคัดกรองตัวแปรด้วยวิธี Lasso, Adaptive Lasso, En และ SCAD โดยเปรียบเทียบกับจำนวนสัมประสิทธิ์ของตัวแบบที่แท้จริง ที่ขนาดของสัมประสิทธิ์ไม่เท่ากับ 0 มีขนาดใหญ่ $(1 < |\beta| < 10)$ พบว่า

1.) ที่จำนวนตัวแปรอิสระ (p) เท่ากับ 20
   - เมื่อ $|S|$ มี 1 และ 2 ค่า ในทุกระดับความสัมพันธ์ของตัวแปรอิสระ การคัดกรองตัวแปรด้วยวิธี Adaptive Lasso สามารถหาค่า $|\hat{S}|$ เข้าใกล้ $|S|$ มากที่สุด ดังนั้นในกรณีนี้การใช้วิธี Adaptive Lasso เพื่อคัดกรองตัวแปรจึงเป็นวิธีที่เหมาะสมที่สุด
   - เมื่อ $|S|$ มี 5 ค่า เมื่อตัวแปรอิสระมีค่าความสัมพันธ์ที่ $\rho = 0$ และ 0.5 การคัดกรองตัวแปรด้วย 4 วิธีข้างต้นหาค่า $|\hat{S}|$ ได้เท่ากับ 0 เท่ากันทุกวิธี นั่นแสดงว่าไม่มีวิธีคัดกรองตัวแปรที่ให้สัมประสิทธิ์ตัวใดไม่เท่ากับ 0 จากการทดสอบสมมติฐาน แต่เมื่อตัวแปรอิสระมีความสัมพันธ์กันที่ $\rho = 0.9$ การคัดกรองตัวแปรด้วยวิธี SCAD สามารถหาค่า $|\hat{S}|$ เข้าใกล้ $|S|$ มากที่สุด ดังนั้นในกรณีนี้การใช้วิธี SCAD เพื่อคัดกรองตัวแปรจึงเป็นวิธีที่เหมาะสมที่สุด

2.) ที่จำนวนตัวแปรอิสระ (p) เท่ากับ 50
   - เมื่อ $|S|$ มี 1 และ 2 ค่า ในทุกระดับความสัมพันธ์ของตัวแปรอิสระ การคัดกรองตัวแปรด้วยวิธี Adaptive Lasso สามารถหาค่า $|\hat{S}|$ และเข้าใกล้ $|S|$ มากที่สุด ดังนั้นในกรณีนี้การใช้วิธี Adaptive Lasso เพื่อคัดกรองตัวแปรจึงเป็นวิธีที่เหมาะสมที่สุด
   - เมื่อ $|S|$ มี 5 ค่า ในทุกระดับความสัมพันธ์ของตัวแปรอิสระ การคัดกรองตัวแปรด้วย 4 วิธีข้างต้นหาค่า $|\hat{S}|$ ได้เท่ากับ 0 เท่ากันทุกวิธี นั่นแสดงว่าไม่มีวิธีคัดกรองตัวแปรที่ให้ค่าสัมประสิทธิ์ตัวใดไม่เท่ากับ 0 จากการทดสอบสมมติฐาน

3.) ที่จำนวนตัวแปรอิสระ (p) เท่ากับ 100
   - เมื่อ $|S|$ 1 ค่า ในทุกระดับความสัมพันธ์ของตัวแปรอิสระ การคัดกรองตัวแปรด้วยวิธี Adaptive Lasso สามารถหาค่า $|\hat{S}|$ ได้เข้าใกล้ $|S|$ มากที่สุด ดังนั้นในกรณีนี้การใช้วิธี Adaptive Lasso เพื่อคัดกรองตัวแปรจึงเป็นวิธีที่เหมาะสมที่สุด
   - เมื่อ $|S|$ มี 2 ค่า เมื่อตัวแปรอิสระมีค่าความสัมพันธ์ที่ $\rho = 0$ การคัดกรองตัวแปรด้วยวิธี Lasso และ Adaptive Lasso สามารถหาค่า $|\hat{S}|$ ได้มากที่สุดและเข้าใกล้ $|S|$ มากที่สุด ดังนั้นในกรณีนี้การใช้วิธี Lasso และวิธี Adaptive Lasso เพื่อคัดกรองตัวแปรจึงเป็นวิธีที่เหมาะสมที่สุด
   - เมื่อ $|S|$ มี 5 ค่า เมื่อตัวแปรอิสระมีค่าความสัมพันธ์ที่ $\rho = 0$ และ 0.5 การคัดกรองตัวแปรด้วย 4 วิธีข้างต้นหาค่า $|\hat{S}|$ ได้เท่ากับ 0 เท่ากันทุกวิธี นั่นแสดงว่าไม่มีวิธีคัดกรองตัวแปรที่ให้ค่าสัมประสิทธิ์ตัวใดไม่เท่ากับ 0 จากการทดสอบสมมติฐาน แต่เมื่อตัวแปรอิสระมีความสัมพันธ์กันที่ $\rho = 0.9$ การคัดกรองตัวแปรด้วยวิธี Adaptive Lasso และวิธี SCAD สามารถหาค่า $|\hat{S}|$ ได้ใกล้เคียงกับ $|S|$ มากที่สุด ดังนั้นในกรณีนี้การใช้วิธี Adaptive Lasso และ SCAD เพื่อคัดกรองตัวแปรจึงเป็นวิธีที่เหมาะสมที่สุด

และจากผลในตารางที่ 4.1.2 ยังสามารถสรุปได้อีกว่า

- ในกรณีที่ p มีจำนวนเพิ่มขึ้นจะทำให้ประสิทธิภาพในการหาค่า $|\hat{S}|$ ให้ใกล้เคียงกับ $|S|$ ลดลง
- ในกรณีที่ ρ มีค่าเพิ่มขึ้นจะทำให้ประสิทธิภาพในการหาค่า $|\hat{S}|$ ให้ใกล้เคียงกับ $|S|$ มากขึ้น
- ในกรณีที่ $|S|$ มีจำนวนเพิ่มขึ้นจะทำให้ประสิทธิภาพในการหาค่า $|\hat{S}|$ ให้ใกล้เคียงกับ $|S|$ ลดลง

และจากตารางที่ 4.1.1 และตารางที่ 4.1.2 จะได้ว่าขนาดของสัมประสิทธิ์ที่ไม่เท่ากับ 0 ที่ขนาดใหญ่ (Large effect Size) จะหาค่า $|\hat{S}|$ ให้ใกล้เคียงกับ $|S|$ ได้ดีกว่าขนาดเล็ก (Small effect Size)

ตารางที่ 4.1.3 แสดงค่าเฉลี่ย(ค่าเบี่ยงเบนมาตรฐาน) ของ FP เมื่อควบคุม FDR ที่ระดับ 0.1 โดยคำนวณจากข้อมูล 100 ชุด กรณีที่ขนาดตัวอย่าง (n) เท่ากับ 10 และขนาดของสัมประสิทธิ์ที่ไม่เท่ากับ 0 มีขนาดเล็ก (Small effect Size)

n : p	|S|	ขนาด(Effect size) ของสัมประสิทธิ์ที่ไม่เท่ากับ 0 มีขนาดเล็ก $(0 < \|\beta\| < 1)$								
		ความสัมพันธ์ (Correlation) ของตัวแปรอิสระ								
		$\rho = 0$			$\rho = 0.5$			$\rho = 0.9$		
		Lasso	Adaptive Lasso	SCAD	Lasso	Adaptive Lasso	SCAD	Lasso	Adaptive Lasso	SCAD
10:20	1	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)
	2	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0.010 (0.100)	0.010 (0.100)
	5	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)
10:50	1	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)
	2	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)
	5	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)
10:100	1	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)
	2	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)
	5	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)

หมายเหตุ ช่องที่ระบายสี หมายถึง วิธีที่เหมาะสมที่สุดในแต่ละกรณี

จากตารางที่ 4.1.3 ซึ่งแสดงผลของ False Positive (FP) โดยเฉลี่ยของข้อมูลจำลองขนาด 10 ระหว่างการคัดกรองตัวแปรด้วยวิธี Lasso, Adaptive Lasso, En และ SCAD โดยเปรียบเทียบกับจำนวนสัมประสิทธิ์ของตัวแบบที่แท้จริง ที่ขนาดของสัมประสิทธิ์ไม่เท่ากับ 0 มีขนาดเล็ก $(0 < |\beta| < 1)$ พบว่า

1. ที่จำนวนตัวแปรอิสระ (p) เท่ากับ 20
    - เมื่อ $|S|$ มี 1 และ 5 ค่า ทุกระดับความสัมพันธ์(Correlation) ของตัวแปรอิสระ การคัดกรองตัวแปรด้วย 4 วิธีข้างต้น มีค่า FP เท่ากับ 0 ทั้งหมด
    - เมื่อ $|S|$ มี 2 ค่า เมื่อตัวแปรอิสระมีค่าความสัมพันธ์ที่ $\rho = 0$ และ 0.5 การคัดกรองตัวแปรด้วย 4 วิธีข้างต้น จะมีค่า FP เท่ากับ 0 แต่เมื่อตัวแปรอิสระมีความสัมพันธ์กันที่ $\rho = 0.9$ การคัดกรองตัวแปรด้วย Lasso มีค่าเท่ากับ 0 และเป็นค่าที่ต่ำที่สุดและเข้าใกล้ 0 มากที่สุด จึงถือได้ว่าสำหรับค่า FP ในกรณีนี้การคัดกรองตัวแปรด้วยวิธี Lasso จึงเป็นวิธีที่เหมาะสมที่สุด
2. ที่จำนวนตัวแปรอิสระ (p) เท่ากับ 50 และ 100
    - เมื่อ $|S|$ มี 1, 2 และ 5 ค่า ตัวแปรอิสระมีค่าความสัมพันธ์ที่ $\rho = 0, 0.5$ และ 0.9 การประมาณค่าด้วย 4 วิธีข้างต้น จะให้ค่า FP เท่ากับ 0 ทั้งหมด

และจากผลในตารางที่ 4.1.3 ยังสามารถสรุปได้อีกว่า

- จำนวนของ p ไม่มีผลต่อประสิทธิภาพในการหาค่า FP ให้เข้าใกล้ 0
- ขนาดของ ρ ไม่มีผลต่อประสิทธิภาพในการหาค่า FP ให้เข้าใกล้ 0
- ขนาดของ $|S|$ มีไม่มีผลต่อประสิทธิภาพในการหาค่า FP ให้เข้าใกล้ 0

<u>ตารางที่ 4.1.4</u> แสดงค่าเฉลี่ย(ค่าเบี่ยงเบนมาตรฐาน) ของ FP เมื่อควบคุม FDR ที่ระดับ 0.1 โดยคำนวณจากข้อมูล 100 ชุด กรณีที่ขนาดตัวอย่าง (n) เท่ากับ 10 และขนาดของสัมประสิทธิ์ที่ไม่เท่ากับ 0 มีขนาดใหญ่ (Large effect Size)

n : p	|S|	ขนาด(Effect size) ของสัมประสิทธิ์ที่ไม่เท่ากับ 0 มีขนาดใหญ่ $(1 < \|\beta\| < 10)$								
		ความสัมพันธ์ (Correlation) ของตัวแปรอิสระ								
		$\rho = 0$			$\rho = 0.5$			$\rho = 0.9$		
		Lasso	Adaptive Lasso	SCAD	Lasso	Adaptive Lasso	SCAD	Lasso	Adaptive Lasso	SCAD
10:20	1	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0.010 (0.100)	0.010 (0.100)
	2	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0.010 (0.100)	0.010 (0.100)	0.010 (0.100)
	5	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)
10:50	1	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)
	2	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)
	5	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)
10:100	1	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)
	2	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)
	5	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0.010 (0.100)	0.010 (0.100)

<u>หมายเหตุ</u> ช่องที่ระบายสี หมายถึง วิธีที่เหมาะสมที่สุดในแต่ละกรณี

จากตารางที่ 4.1.4 ซึ่งแสดงผลของ False Positive โดยเฉลี่ยของข้อมูลจำลองขนาด 10 ระหว่างการคัดกรองตัวแปรด้วยวิธี Lasso, Adaptive Lasso, En และ SCAD โดยเปรียบเทียบกับจำนวนสัมประสิทธิ์ของตัวแบบที่แท้จริง ที่ขนาดของสัมประสิทธิ์ไม่เท่ากับ 0 มีขนาดใหญ่ $(1 < |\beta| < 10)$ พบว่า

1. ที่จำนวนตัวแปรอิสระ (p) เท่ากับ 20
   - เมื่อ $|S|$ มี 1 ค่า ที่ระดับความสัมพันธ์ที่ $\rho = 0$ และ 0.5 การคัดกรองตัวแปรด้วย 4 วิธีข้างต้น มีค่า FP เท่ากับ 0 ทั้งหมด แต่ที่ $\rho = 0.9$ การคัดกรองตัวแปรด้วยวิธี Lasso มีค่า FP เท่ากับ 0 ซึ่งเป็นค่าที่ต่ำที่สุดและเข้าใกล้ 0 มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วยวิธี Lasso จึงเป็นวิธีที่เหมาะสมที่สุด
   - เมื่อ $|S|$ มี 2 และ 5 ค่า ทุกระดับความสัมพันธ์(Correlation) ของตัวแปรอิสระ การคัดกรองตัวแปรด้วย 4 วิธีข้างต้น มีค่า FP เท่ากับ 0 ทั้งหมด
2. ที่จำนวนตัวแปรอิสระ (p) เท่ากับ 50
   - เมื่อ $|S|$ มี 1, 2 และ 5 ค่า ตัวแปรอิสระมีค่าความสัมพันธ์ที่ $\rho = 0$, 0.5 และ 0.9 การคัดกรองตัวแปรด้วย 4 วิธีข้างต้น จะให้ค่า FP เท่ากับ 0 ทั้งหมด
3. ที่จำนวนตัวแปรอิสระ (p) เท่ากับ 100
   - เมื่อ $|S|$ มี 1 และ 2 ค่า ทุกระดับความสัมพันธ์(Correlation) ของตัวแปรอิสระ การคัดกรองตัวแปรด้วย 4 วิธีข้างต้น มีค่า FP เท่ากับ 0 ทั้งหมด
   - เมื่อ $|S|$ มี 5 ค่า ที่ระดับความสัมพันธ์(Correlation) ของตัวแปรอิสระ $\rho = 0$ และ 0.5 การคัดกรองตัวแปรด้วย 4 วิธีข้างต้น มีค่า FP เท่ากับ 0 ทั้งหมด แต่ที่ $\rho = 0.9$ การคัดกรองตัวแปรด้วยวิธี Lasso มีค่า FP เท่ากับ 0 ซึ่งเป็นค่าที่ต่ำที่สุดและเข้าใกล้ 0 มากที่สุด ดังนั้น ในกรณีนี้การคัดกรองตัวแปรด้วยวิธี Lasso จึงเป็นวิธีที่เหมาะสมที่สุด

และจากผลในตารางที่ 4.1.4 ยังสามารถสรุปได้อีกว่า

- จำนวนของ p ไม่มีผลต่อประสิทธิภาพในการหาค่า FP ให้เข้าใกล้ 0
- ขนาดของ ρ ไม่มีผลต่อประสิทธิภาพในการหาค่า FP ให้เข้าใกล้ 0
- ขนาดของ $|S|$ มีไม่มีผลต่อประสิทธิภาพในการหาค่า FP ให้เข้าใกล้

และจากตารางที่ 4.1.3 และตารางที่ 4.1.4 จะได้ว่าขนาดของสัมประสิทธิ์ที่ไม่เท่ากับ 0 ไม่มีผลต่อประสิทธิภาพในการหาค่า FP ให้เข้าใกล้ 0

ตารางที่ 4.1.5 แสดงค่าเฉลี่ย(ค่าเบี่ยงเบนมาตรฐาน) ของ FN เมื่อควบคุม FDR ที่ระดับ 0.1 โดยคำนวณจากข้อมูล 100 ชุด กรณีที่ขนาดตัวอย่าง (n) เท่ากับ 10 และขนาดของสัมประสิทธิ์ที่ไม่เท่ากับ 0 มีขนาดเล็ก (Small effect Size)

n : p	|S|	ขนาด(Effect size) ของสัมประสิทธิ์ที่ไม่เท่ากับ 0 มีขนาดเล็ก $(0 < \|\beta\| < 1)$								
		ความสัมพันธ์ (Correlation) ของตัวแปรอิสระ								
		$\rho = 0$			$\rho = 0.5$			$\rho = 0.9$		
		Lasso	Adaptive Lasso	SCAD	Lasso	Adaptive Lasso	SCAD	Lasso	Adaptive Lasso	SCAD
10:20	1	1 (0.000)	1 (0.000)	1 (0.000)	1 (0.000)	1 (0.000)	1 (0.000)	1 (0.000)	1 (0.000)	1 (0.000)
	2	2 (0.000)	2 (0.000)	2 (0.000)	2 (0.000)	2 (0.000)	2 (0.000)	2 (0.000)	2 (0.000)	2 (0.000)
	5	5 (0.000)	5 (0.000)	5 (0.000)	5 (0.000)	5 (0.000)	5 (0.000)	5 (0.000)	4.990 (0.100)	5 (0.000)
10:50	1	1 (0.000)	1 (0.000)	1 (0.000)	1 (0.000)	1 (0.000)	1 (0.000)	1 (0.000)	1 (0.000)	1 (0.000)
	2	2 (0.000)	2 (0.000)	2 (0.000)	2 (0.000)	2 (0.000)	2 (0.000)	2 (0.000)	2 (0.000)	2 (0.000)
	5	5 (0.000)	5 (0.000)	5 (0.000)	5 (0.000)	5 (0.000)	5 (0.000)	5 (0.000)	5 (0.000)	5 (0.000)
10:100	1	1 (0.000)	1 (0.000)	1 (0.000)	1 (0.000)	1 (0.000)	1 (0.000)	1 (0.000)	1 (0.000)	1 (0.000)
	2	2 (0.000)	2 (0.000)	2 (0.000)	2 (0.000)	2 (0.000)	2 (0.000)	2 (0.000)	2 (0.000)	2 (0.000)
	5	5 (0.000)	5 (0.000)	5 (0.000)	5 (0.000)	5 (0.000)	5 (0.000)	5 (0.000)	5 (0.000)	5 (0.000)

หมายเหตุ ช่องที่ระบายสี หมายถึง วิธีที่เหมาะสมที่สุดในแต่ละกรณี

จากตารางที่ 4.1.5 ซึ่งแสดงผลของ False Negative โดยเฉลี่ยของข้อมูลจำลองขนาด 10 ระหว่างการคัดกรองตัวแปรด้วยวิธี Lasso, Adaptive Lasso, En และ SCAD โดยเปรียบเทียบกับจำนวนสัมประสิทธิ์ของตัวแบบที่แท้จริง ที่ขนาดของสัมประสิทธิ์ไม่เท่ากับ 0 มีขนาดเล็ก $(0 < |\beta| < 1)$ พบว่า

1. ที่จำนวนตัวแปรอิสระ (p) เท่ากับ 20
   - เมื่อ $|S|$ มี 5 ค่า มี 1 และ 2 ค่า ทุกระดับความสัมพันธ์(Correlation) ของตัวแปรอิสระ การคัดกรองตัวแปรด้วย 4 วิธีข้างต้น มีค่า FN เท่ากับ 0 ทั้งหมด
   - เมื่อ $|S|$ มี 5 ค่า มี 5 ค่า ที่ระดับความสัมพันธ์(Correlation) ของตัวแปรอิสระ $\rho = 0$ และ 0.5 การคัดกรองตัวแปรด้วย 4 วิธีข้างต้น มีค่า FP เท่ากับ 0 ทั้งหมด แต่ที่ $\rho = 0.9$ การคัดกรองตัวแปรด้วยวิธี Lasso มีค่า FP ต่ำที่สุดและเข้าใกล้ 0 มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วยวิธี Lasso จึงเป็นวิธีที่เหมาะสมที่สุด
2. ที่จำนวนตัวแปรอิสระ (p) เท่ากับ 50 และ 100
   - ในทุกกรณีที่ตัวแปรอิสระมีค่าความสัมพันธ์กันที่ $\rho = 0$, 0.5 และ 0.9 การคัดกรองตัวแปรด้วย 4 วิธีข้างต้น จะให้ค่า FN เท่ากับจำนวนค่าสัมประสิทธิ์ของตัวแบบที่แท้จริงไม่เท่ากับ 0 ของแต่ละกรณี

และจากผลในตารางที่ 4.1.5 ยังสามารถสรุปได้อีกว่า

- จำนวนของ p ไม่มีผลต่อประสิทธิภาพในการหาค่า FP ให้เข้าใกล้ 0
- ขนาดของ ρ ไม่มีผลต่อประสิทธิภาพในการหาค่า FP ให้เข้าใกล้ 0
- ขนาดของ $|S|$ มีไม่มีผลต่อประสิทธิภาพในการหาค่า FP ให้เข้าใกล้ 0

<u>ตารางที่ 4.1.6</u> แสดงค่าเฉลี่ย(ค่าเบี่ยงเบนมาตรฐาน) ของ FN เมื่อควบคุม FDR ที่ระดับ 0.1 โดยคำนวณจากข้อมูล 100 ชุด กรณีที่ขนาดตัวอย่าง (n) เท่ากับ 10 และขนาดของสัมประสิทธิ์ที่ไม่เท่ากับ 0 มีขนาดใหญ่ (Large effect Size)

n : p	|S|	ขนาด(Effect size) ของสัมประสิทธิ์ที่ไม่เท่ากับ 0 มีขนาดใหญ่ $(1 < \|\beta\| < 10)$								
		ความสัมพันธ์ (Correlation) ของตัวแปรอิสระ								
		$\rho = 0$			$\rho = 0.5$			$\rho = 0.9$		
		Lasso	Adaptive Lasso	SCAD	Lasso	Adaptive Lasso	SCAD	Lasso	Adaptive Lasso	SCAD
10:20	1	0.550	0.520	0.580	0.470	0.420	0.540	0.590	0.420	0.470
		(0.500)	(0.502)	(0.486)	(0.502)	(0.496)	(0.501)	(0.494)	(0.496)	(0.501)
	2	1.920	1.900	1.930	1.970	1.930	1.950	1.870	1.790	1.870
		(0.273)	(0.302)	(0.256)	(0.171)	(0.293)	(0.219)	(0.338)	(0.409)	(0.338)
	5	5.000	5.000	5.000	5.000	5.000	5.000	5.000	5.000	4.990
		(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.100)
10:50	1	0.440	0.430	0.540	0.600	0.580	0.610	0.680	0.610	0.660
		(0.498)	(0.497)	(0.501)	(0.492)	(0.496)	(0.490)	(0.469)	(0.490)	(0.476)
	2	1.980	1.930	1.970	2.000	1.980	2.000	1.930	1.910	1.930
		(0.141)	(0.256)	(0.171)	(0.000)	(0.141)	(0.000)	(0.256)	(0.287)	(0.256)
	5	5.000	5.000	5.000	5.000	5.000	5.000	5.000	5.000	5.000
		(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
10:100	1	0.700	0.660	0.750	0.640	0.580	0.660	0.670	0.650	0.710
		(0.461)	(0.476)	(0.435)	(0.482)	(0.496)	(0.476)	(0.472)	(0.479)	(0.456)
	2	1.980	1.980	1.990	1.990	1.980	1.980	1.960	1.950	1.980
		(0.141)	(0.141)	(0.100)	(0.100)	(0.141)	(0.141)	(0.196)	(0.219)	(0.141)
	5	5.000	5.000	5.000	5.000	5.000	5.000	5.000	5.000	5.000
		(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)

หมายเหต ช่องที่ระบายสี หมายถึง วิธีที่เหมาะสมที่สดในแต่ละกรณี

จากตารางที่ 4.1.6 ซึ่งแสดงผลของ False Negative โดยเฉลี่ยของข้อมูลจำลองขนาด 10 ระหว่างการคัดกรองตัวแปรด้วยวิธี Lasso, Adaptive Lasso, En และ SCAD โดยเปรียบเทียบกับจำนวนสัมประสิทธิ์ของตัวแบบที่แท้จริง ที่ขนาดของสัมประสิทธิ์ไม่เท่ากับ 0 มีขนาดใหญ่ $(1 < |\beta| < 10)$ พบว่า

1.) ที่จำนวนตัวแปรอิสระ (p) เท่ากับ 20
   - เมื่อ $|S|$ มี 1 และ 2 ค่า ในทุกระดับความสัมพันธ์ของตัวแปรอิสระ การคัดกรองตัวแปรด้วยวิธี Adaptive Lasso จะให้ค่า FN ได้ต่ำที่สุดหรือเข้าใกล้ 0 มากที่สุด ดังนั้นในกรณีนี้การใช้วิธี Adaptive Lasso เพื่อคัดกรองตัวแปรจึงเป็นวิธีที่เหมาะสมที่สุด
   - เมื่อ $|S|$ มี 5 ค่า เมื่อตัวแปรอิสระมีค่าความสัมพันธ์ที่ $\rho = 0$ และ 0.5 การคัดกรองตัวแปรด้วย 4 วิธีข้างต้นจะได้ค่า FN เท่ากันทั้งหมด แต่เมื่อตัวแปรอิสระมีความสัมพันธ์กันที่ $\rho = 0.9$ การคัดกรองตัวแปรด้วยวิธี SCAD จะได้ FNต่ำที่สุดหรือเข้าใกล้ 0 มากที่สุด ดังนั้นในกรณีนี้การใช้วิธี SCAD เพื่อคัดกรองตัวแปรจึงเป็นวิธีที่เหมาะสมที่สุด

2.) ที่จำนวนตัวแปรอิสระ (p) เท่ากับ 50
   - เมื่อ $|S|$ มี 1 และ 2 ค่า ในทุกระดับความสัมพันธ์ของตัวแปรอิสระ การคัดกรองตัวแปรด้วยวิธี Adaptive Lasso จะให้ค่า FN ได้ต่ำที่สุดหรือเข้าใกล้ 0 มากที่สุด ดังนั้นในกรณีนี้การใช้วิธี Adaptive Lasso เพื่อคัดกรองตัวแปรจึงเป็นวิธีที่เหมาะสมที่สุด
   - เมื่อ $|S|$ มี 5 ค่า ที่ตัวแปรอิสระมีค่าความสัมพันธ์กันที่ $\rho = 0, 0.5$ และ 0.9 การคัดกรองตัวแปรด้วย 4 วิธีข้างต้น จะให้ค่า FN เท่ากันทั้งหมด

3.) ที่จำนวนตัวแปรอิสระ (p) เท่ากับ 100
   - เมื่อ $|S|$ มี 1 ค่า ในทุกระดับความสัมพันธ์ของตัวแปรอิสระ การคัดกรองตัวแปรด้วยวิธี Adaptive Lasso จะให้ค่า FN ได้ต่ำที่สุดหรือเข้าใกล้ 0 มากที่สุด ดังนั้นในกรณีนี้การใช้วิธี Adaptive Lasso เพื่อคัดกรองตัวแปรจึงเป็นวิธีที่เหมาะสมที่สุด
   - เมื่อ $|S|$ มี 2 ค่า เมื่อตัวแปรอิสระมีความสัมพันธ์กันที่ $\rho = 0$ และ 0.5 การคัดกรองตัวแปรด้วยวิธี Lasso และ Adaptive Lasso จะให้ค่า FN ได้ต่ำที่สุดหรือเข้าใกล้ 0 มากที่สุด ดังนั้นในกรณีนี้การใช้วิธี Lasso และ Adaptive Lasso เพื่อคัดกรองตัวแปรจึงเป็นวิธีที่เหมาะสมที่สุด
   - ในทุกกรณีที่ตัวแปรอิสระมีค่าความสัมพันธ์กันที่ $\rho = 0, 0.5$ และ 0.9 การคัดกรองตัวแปรด้วย 4 วิธีข้างต้น จะให้ค่า FN เท่ากันทั้งหมด

และจากผลในตารางที่ 4.1.5 ยังสามารถสรุปได้อีกว่า
- ในกรณีที่ p มีจำนวนเพิ่มขึ้นจะทำให้ประสิทธิภาพในการหาค่า FN ให้เข้าใกล้ 0 ลดลง
- ในกรณีที่ ρ มีค่าเพิ่มขึ้นจะทำให้ประสิทธิภาพในการหาค่า FN ให้เข้าใกล้ 0 มากขึ้น

- ในกรณีที่ $|S|$ มีจำนวนเพิ่มขึ้นจะทำให้ประสิทธิภาพในการหาค่า FN ให้เข้าใกล้ 0 ลดลง

และจากตารางที่ 4.1.5 และตารางที่ 4.1.6 จะได้ว่าขนาดของสัมประสิทธิ์ที่ไม่เท่ากับ 0 ที่ขนาดใหญ่ (Large effect Size) จะหาค่า FN เข้าใกล้ 0 ได้ดีกว่าขนาดเล็ก (Small effect Size)

จากตารางที่ 4.1.1 – 4.1.6 เมื่อพิจารณาจากค่า $|\hat{S}|$, FP และ FN จะได้ว่า เมื่อขนาดตัวอย่าง (n) เท่ากับ 10 ค่าของ $|\hat{S}|$ และ FN จะไปในทิศทางเดียวกัน นั่นคือการคัดกรองตัวแปรด้วยวิธี Adaptive Lasso จะเหมาะสมที่สุด แต่ค่าของ FP จะได้วิธี Lasso เป็นวิธีที่เหมาะสมที่สุด ดังนั้นจะได้ว่าการคัดกรองตัวแปรด้วยวิธี Adaptive Lasso ในกรณีที่ n = 10 จะเหมาะสมและมีอำนาจการทดสอบมากที่สุด

4.2 ผลการเปรียบเทียบค่าความผิดพลาดในการตรวจจับเชิงบวก (FP) ค่าความผิดพลาดในการตรวจจับเชิงลบ (FN) และจำนวนของสัมประสิทธิ์ของตัวแปรอิสระที่มีค่าไม่เท่ากับ 0 จากการทดสอบสมมติฐานโดยเฉลี่ย เมื่อควบคุม FDR ที่ระดับ 0.1 ของข้อมูลจำลองขนาด 100 ระหว่างการคัดกรองตัวแปรด้วยวิธี Lasso, Adaptive Lasso, EN และ SCAD

ในส่วนนี้ผู้วิจัยต้องการศึกษาเปรียบเทียบการคัดกรองตัวแปรจากวิธี Lasso, Adaptive Lasso, EN และ SCAD และเพื่อพิจารณาว่าปัจจัยใดที่ส่งผลต่อประสิทธิภาพการทำงานของวิธีการคัดกรองตัวแปรแต่ละวิธี ภายใต้ปัจจัย ดังต่อไปนี้

1. อัตราส่วนระหว่างขนาดตัวอย่างและจำนวนตัวแปรอิสระ คือ 1:2, 1:5 และ 1:10
2. ร้อยละของจำนวนสัมประสิทธิ์ที่ไม่เท่ากับ 0 $(\beta \neq 0)$ เมื่อเทียบกับขนาดตัวอย่าง คือ ร้อยละ 10, 20 และ 50
3. ขนาด (Effect Size) ของสัมประสิทธิ์ที่ไม่เท่ากับ 0 คือ Small Size $(0 < |\beta| < 1)$ และ Large Size $(1 < |\beta| < 10)$
4. ความสัมพันธ์ (Correlation) ของตัวแปรอิสระ คือ ρ = 0, 0.5, 0.9

โดยแสดงผลในตารางที่ 4.2.1 - 4.2.6 โดยแต่ละตารางมีรายละเอียดดังนี้

เกณฑ์ที่ใช้ในการวัด	ตารางที่	ปัจจัยที่ใช้ในการพิจารณา	วิธีการคัดกรองตัวแปรที่ต้องการเปรียบเทียบ
$\|\hat{S}\|$	4.2.1	• n ; p • จำนวนสัมประสิทธิ์ของโมเดลที่แท้จริงที่มีค่าไม่เท่ากับ 0 ($\beta \neq 0$) • ความสัมพันธ์ (Correlation) ของตัวแปรอิสระ • ขนาด(Effect Size) ของสัมประสิทธิ์ที่ไม่เท่ากับ 0	1. Lasso 2. Adaptive Lasso 3. EN 4. SCAD
	4.2.2		
FP	4.2.3		
	4.2.4		
FN	4.2.5		
	4.2.6		

ตารางที่ 4.2.1 แสดงค่าเฉลี่ย(ค่าเบี่ยงเบนมาตรฐาน) ของ $|S|$ เมื่อควบคุม FDR ที่ระดับ 0.1 โดยคำนวณจากข้อมูล 100 ชุด กรณีที่ขนาดตัวอย่าง (n) เท่ากับ 100 และขนาดของสัมประสิทธิ์ที่ไม่เท่ากับ 0 มีขนาดเล็ก (Small effect Size)

n : p	$\|S\|$	ขนาด(Effect size) ของสัมประสิทธิ์ที่ไม่เท่ากับ 0 เป็น Small Size $(0 < \|\beta\| < 1)$											
		ความสัมพันธ์ (Correlation) ของตัวแปรอิสระ											
		$\rho = 0$				$\rho = 0.5$				$\rho = 0.9$			
		Lasso	Adaptive Lasso	EN	SCAD	Lasso	Adaptive Lasso	EN	SCAD	Lasso	Adaptive Lasso	EN	SCAD
100:200	10	1.910	2.870	0.190	2.880	1.690	2.930	0.210	2.880	0.250	2.500	0	2.640
		(1.443)	(1.612)	(0.465)	(1.546)	(1.383)	(1.671)	(0.574)	(1.683)	(0.500)	(1.823)	(0.000)	(2.153)
	20	0.300	1.150	0.010	0.960	0.230	1.000	0.010	0.820	0.040	2.080	0	2.330
		(0.704)	(1.766)	(0.100)	(1.325)	(0.679)	(1.524)	(0.100)	(1.250)	(0.197)	(1.998)	(0.000)	(2.207)
	50	0	0.010	0	0.010	0	0.070	0	0	0.010	0.860	0	1.820
		(0.000)	(0.100)	(0.000)	(0.100)	(0.000)	(0.256)	(0.000)	(0.000)	(0.100)	(1.155)	(0.000)	(2.110)
100:500	10	0	1.930	0.080	1.840	1.390	2.020	0.060	1.930	0.310	2.220	0.010	1.640
		(0.000)	(1.472)	(0.307)	(1.419)	(1.238)	(1.706)	(0.239)	(1.591)	(0.706)	(1.988)	(0.100)	(1.709)
	20	0.160	0.290	0.020	0.300	0.120	0.240	0	0.300	0.060	0.930	0	0.730
		(0.465)	(0.640)	(0.141)	(0.611)	(0.433)	(0.683)	(0.000)	(0.704)	(0.239)	(1.365)	(0.000)	(1.109)
	50	0.020	0.020	0	0.030	0.010	0.030	0	0	0	0.180	0	0.130
		(0.141)	(0.141)	(0.000)	(0.171)	(0.100)	(0.223)	(0.000)	(0.000)	(0.000)	(0.435)	(0.000)	(0.442)
100:1000	10	0.910	1.430	0.040	1.320	0.860	1.240	0.030	1.280	0.330	1.850	0	1.150
		(1.120)	(1.416)	(0.243)	(1.392)	(1.083)	(1.372)	(0.171)	(1.334)	(0.651)	(2.002)	(0.000)	(1.424)
	20	0.100	0.270	0.010	0.260	0.050	0.190	0	0.130	0.030	0.410	0	0.180
		(0.362)	(0.709)	(0.100)	(0.645)	(0.219)	(0.563)	(0.000)	(0.393)	(0.171)	(0.922)	(0.000)	(0.435)
	50	0	0	0	0	0.010	0.010	0	0.010	0	0.170	0	0.100
		(0.000)	(0.000)	(0.000)	(0.000)	(0.100)	(0.100)	(0.000)	(0.100)	(0.000)	(0.570)	(0.000)	(0.389)

หมายเหตุ ช่องที่ระบายสี หมายถึง วิธีที่เหมาะสมที่สุดในแต่ละกรณี

จากตารางที่ 4.2.1 ซึ่งแสดงผลของ $|\hat{S}|$ โดยเฉลี่ยของข้อมูลจำลองขนาด 100 ระหว่างการคัดกรองตัวแปรด้วยวิธี Lasso, Adaptive Lasso, En และ SCAD โดยเปรียบเทียบกับจำนวนสัมประสิทธิ์ของตัวแบบที่แท้จริง ที่ขนาดของสัมประสิทธิ์ไม่เท่ากับ 0 มีขนาดเล็ก $(0 < |\beta| < 1)$ พบว่า

1.) ที่จำนวนตัวแปรอิสระ (p) เท่ากับ 200

- เมื่อ $|S|$ มี 10 ค่า เมื่อตัวแปรอิสระมีค่าความสัมพันธ์ที่ $\rho = 0$ และ 0.9 การคัดกรองตัวแปรด้วยวิธี SCAD สามารถหาค่า $|\hat{S}|$ ได้มากที่สุดและเข้าใกล้ $|S|$ มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วยวิธี SCAD จึงมีความเหมาะสมที่สุด และที่ $\rho = 0.5$ การคัดกรองตัวแปรด้วยวิธี Adaptive Lasso จะมีค่ามากที่สุดและเข้าใกล้ $|S|$ มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วยวิธี Adaptive Lasso จึงเหมาะสมที่สุด
- เมื่อ $|S|$ มี 20 ค่า เมื่อตัวแปรอิสระมีค่าความสัมพันธ์ที่ $\rho = 0$ และ 0.5 การคัดกรองตัวแปรด้วยวิธี Adaptive Lasso สามารถหาค่า $|\hat{S}|$ ได้มากที่สุดและเข้าใกล้ $|S|$ มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วยวิธี Adaptive Lasso จึงมีความเหมาะสมที่สุด และที่ $\rho = 0.9$ การคัดกรองตัวแปรด้วยวิธี SCAD จะมีค่ามากที่สุดและเข้าใกล้ $|S|$ มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วยวิธี SCAD จึงเหมาะสมที่สุด
- เมื่อ $|S|$ มี 50 ค่า เมื่อตัวแปรอิสระมีค่าความสัมพันธ์ที่ $\rho = 0$ และ 0.5 การคัดกรองตัวแปรด้วยวิธี Adaptive Lasso และ SCAD สามารถหาค่า $|\hat{S}|$ ได้เท่ากันซึ่งมีค่ามากที่สุดและเข้าใกล้ $|S|$ มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วยวิธี Adaptive Lasso และ SCAD จึงมีความเหมาะสมที่สุด และที่ $\rho = 0.5$ การคัดกรองตัวแปรด้วยวิธี Adaptive Lasso จะมีค่ามากที่สุดและเข้าใกล้ $|S|$ มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วยวิธี Adaptive Lasso จึงเหมาะสมที่สุด และที่ $\rho = 0.9$ การคัดกรองตัวแปรด้วยวิธี SCAD จะมีค่ามากที่สุดและเข้าใกล้ $|S|$ มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วยวิธี SCAD จึงเหมาะสมที่สุด

2.) ที่จำนวนตัวแปรอิสระ (p) เท่ากับ 500

- เมื่อ $|S|$ มี 10 ค่า เมื่อตัวแปรอิสระมีค่าความสัมพันธ์ที่ $\rho = 0$ และ 0.9 การคัดกรองตัวแปรด้วยวิธี วิธี Adaptive Lasso สามารถหาค่า $|\hat{S}|$ ได้มากที่สุดและเข้าใกล้ $|S|$ มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วยวิธี วิธี Adaptive Lasso จึงมีความเหมาะสมที่สุด และที่ $\rho = 0.5$ การคัดกรองตัวแปรด้วยวิธี SCAD จะมีค่ามากที่สุดและเข้าใกล้ $|S|$ มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วย SCAD จึงเหมาะสมที่สุด
- เมื่อ $|S|$ มี 20 ค่า เมื่อตัวแปรอิสระมีค่าความสัมพันธ์ที่ $\rho = 0$ และ 0.5 การคัดกรองตัวแปรด้วยวิธี SCAD สามารถหาค่า $|\hat{S}|$ ได้มากที่สุดและเข้าใกล้ $|S|$ มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วยวิธี SCAD จึงมีความเหมาะสมที่สุด และที่ $\rho = 0.9$ การคัดกรองตัวแปรด้วยวิธี Adaptive Lasso จะมีค่ามากที่สุดและเข้าใกล้ $|S|$ มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วยวิธี Adaptive Lasso จึงเหมาะสมที่สุด

- เมื่อ $|S|$ มี 50 ค่า เมื่อตัวแปรอิสระมีค่าความสัมพันธ์ที่ $\rho = 0$ และ 0.9 การคัดกรองตัวแปรด้วยวิธี SCAD สามารถหาค่า $|\hat{S}|$ ได้มากที่สุดและเข้าใกล้ $|S|$ มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วยวิธี SCAD จึงมีความเหมาะสมที่สุด และที่ $\rho = 0.5$ การคัดกรองตัวแปรด้วยวิธี Adaptive Lasso จะมีค่ามากที่สุดและเข้าใกล้ $|S|$ มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วยวิธี Adaptive Lasso จึงเหมาะสมที่สุด

3.) ที่จำนวนตัวแปรอิสระ (p) เท่ากับ 1000

- เมื่อ $|S|$ มี 10 และ 20 ค่า เมื่อตัวแปรอิสระมีค่าความสัมพันธ์ที่ $\rho = 0$ และ 0.9 การคัดกรองตัวแปรด้วยวิธี วิธี Adaptive Lasso สามารถหาค่า $|\hat{S}|$ ได้มากที่สุดและเข้าใกล้ $|S|$ มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วยวิธี วิธี Adaptive Lasso จึงมีความเหมาะสมที่สุด และที่ $\rho = 0.5$ การคัดกรองตัวแปรด้วยวิธี SCAD จะมีค่ามากที่สุดและเข้าใกล้ $|S|$ มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วย SCAD จึงเหมาะสมที่สุด
- เมื่อ $|S|$ มี 50 ค่า เมื่อตัวแปรอิสระมีค่าความสัมพันธ์ที่ $\rho = 0$ การคัดกรองตัวแปรทั้ง 4 วิธี หาค่า $|\hat{S}|$ ได้เท่ากับ 0 เท่ากันทุกวิธี นั่นแสดงว่าไม่มีวิธีคัดกรองตัวแปรใดที่ให้ค่าสัมประสิทธิ์ตัวใดไม่เท่ากับ 0 จากการทดสอบสมมติฐาน และที่ $\rho = 0.5$ การคัดกรองตัวแปรด้วยวิธี Lasso, Adaptive Lasso และ SCAD จะให้ค่า $|\hat{S}|$ เข้าใกล้ $|S|$ มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วยวิธี Lasso, Adaptive Lasso และ SCAD จึงเหมาะสมที่สุด และ ที่ $\rho = 0.9$ การคัดกรองตัวแปรด้วยวิธี Adaptive Lasso จะให้ค่า $|\hat{S}|$ เข้าใกล้ $|S|$ มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วยวิธี Adaptive Lasso จึงเหมาะสมที่สุด

และจากผลในตารางที่ 4.2.1 ยังสามารถสรุปได้อีกว่า

- ในกรณีที่ p มีจำนวนเพิ่มขึ้นจะทำให้ประสิทธิภาพในการหาค่า $|\hat{S}|$ ให้ใกล้เคียงกับ $|S|$ ลดลง
- ในกรณีที่ ρ มีค่าเพิ่มขึ้นจะทำให้ประสิทธิภาพในการหาค่า $|\hat{S}|$ ให้ใกล้เคียงกับ $|S|$ ลดลง
- ในกรณีที่ $|S|$ มีจำนวนเพิ่มขึ้นจะทำให้ประสิทธิภาพในการหาค่า $|\hat{S}|$ ให้ใกล้เคียงกับ $|S|$ ลดลง

ตารางที่ 4.2.2 แสดงค่าเฉลี่ย(ค่าเบี่ยงเบนมาตรฐาน) ของ $|\hat{S}|$ เมื่อควบคุม FDR ที่ระดับ 0.1 โดยคำนวณจากข้อมูล 100 ชุด กรณีที่ขนาดตัวอย่าง (n) เท่ากับ 100 และขนาดของสัมประสิทธิ์ที่ไม่เท่ากับ 0 มีขนาดใหญ่ (Large effect Size)

n : p	$\|S\|$	ขนาด(Effect size) ของสัมประสิทธิ์ที่ไม่เท่ากับ 0 มีขนาดใหญ่ $(1 < \|\beta\| < 10)$											
		ความสัมพันธ์ (Correlation) ของตัวแปรอิสระ											
		$\rho = 0$				$\rho = 0.5$				$\rho = 0.9$			
		Lasso	Adaptive Lasso	EN	SCAD	Lasso	Adaptive Lasso	EN	SCAD	Lasso	Adaptive Lasso	EN	SCAD
100:200	10	9.860	10.000	7.560	10.000	9.640	10.010	0.100	10.000	8.02	10.620	3.730	13.100
		(0.377)	(0.000)	(2.257)	(0.000)	(0.811)	(0.100)	(6.390)	(0.000)	(1.428)	(1.568)	(3.428)	(2.418)
	20	3.050	8.050	0.280	4.040	2.860	8.430	0.020	4.490	0.780	9.910	0	8.740
		(4.205)	(5.391)	(1.248)	(4.494)	(0.000)	(5.040)	(0.141)	(04.491)	(1.851)	(3.911)	(0.000)	(4.113)
	50	0.050	0.060	0	0.080	0	0.010	0	0.010	0	0.390	0	2.420
		(0.261)	(0.239)	(0.000)	(0.272)	(0.000)	(0.100)	(0.000)	(0.100)	(0.000)	(1.034)	(0.000)	(2.606)
100:500	10	9.140	9.590	2.400	10.000	8.260	9.180	1.500	9.990	6.160	11.120	0.080	12.950
		(1.428)	(0.817)	(2.404)	(0.000)	(2.111)	(1.298)	(2.492)	(0.100)	(2.830)	(2.166)	(0.706)	(2.271)
	20	0.180	1.450	0.010	0.910	0.290	1.460	0	0.930	0.040	3.580	0	2.130
		(0.557)	(2.328)	(0.100)	(1.491)	(0.820)	(2.199)	(0.000)	(1.546)	(0197)	(3.462)	(0.000)	(2.452)
	50	0.010	0.020	0	0.020	0.010	0.020	0	0.020	0	0.100	0	0.240
		(0.100)	(0.141)	(0.000)	(0.141)	(0.100)	(0.141)	(0.000)	(0.200)	(0.000)	(0.438)	(0.000)	(0.653)
100:1000	10	6.430	7.760	0.970	9.740	5.800	7.060	0.660	9.530	4.670	10.870	0.040	11.560
		(2.405)	(1.848)	(1.374)	(1.219)	(2.704)	(2.228)	(1.409)	(1.856)	(3.097)	(2.717)	(0.243)	(2.965)
	20	0.110	0.330	0	0.290	0.100	0.470	0.010	0.270	0.060	1.100	0	0.710
		(0.373)	(0.804)	(0.000)	(0.868)	(0.333)	(1.259)	(0.100)	(0.633)	(0.278)	(1.828)	(0.000)	(1.365)
	50	0	0	0	0	0.010	0.020	0	0	0.020	0.030	0	0.020
		(0.000)	(0.000)	(0.000)	(0.000)	(0.100)	(0.141)	(0.000)	(0.000)	(0.141)	(0.223)	(0.000)	(0.141)

หมายเหตุ ช่องที่ระบายสี หมายถึง วิธีที่เหมาะสมที่สุดในแต่ละกรณี

จากตารางที่ 4.2.2 ซึ่งแสดงผลของ $|\hat{S}|$ โดยเฉลี่ยของข้อมูลจำลองขนาด 100 ระหว่างการคัดกรองตัวแปรด้วยวิธี Lasso, Adaptive Lasso, En และ SCAD โดยเปรียบเทียบกับจำนวนสัมประสิทธิ์ของตัวแบบที่แท้จริง ที่ขนาดของสัมประสิทธิ์ไม่เท่ากับ 0 มีขนาดใหญ่ $(1 < |\beta| < 10)$ พบว่า

1.) ที่จำนวนตัวแปรอิสระ (p) เท่ากับ 200

- เมื่อ $|S|$ มี 10 ค่า และตัวแปรอิสระมีค่าความสัมพันธ์ที่ $\rho = 0$ การคัดกรองตัวแปรด้วยวิธี Adaptive Lasso และ SCAD สามารถหาค่า $|\hat{S}|$ ได้มากที่สุดและเข้าใกล้ $|S|$ มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วยวิธี Adaptive Lasso และ SCAD จึงมีความเหมาะสมมากที่สุด และที่ $\rho = 0.5$ และ 0.9 การคัดกรองตัวแปรด้วยวิธี Adaptive Lasso จะมีค่ามากที่สุดและเข้าใกล้ $|S|$ มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วยวิธี Adaptive Lasso จึงเหมาะสมมากที่สุด
- เมื่อ $|S|$ มี 20 ค่า และตัวแปรอิสระมีค่าความสัมพันธ์ที่ $\rho = 0$ การคัดกรองตัวแปรด้วยวิธี Adaptive Lasso สามารถหาค่า $|\hat{S}|$ ได้มากที่สุดและเข้าใกล้ $|S|$ มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วยวิธี Adaptive Lasso จึงมีความเหมาะสมมากที่สุด และที่ $\rho = 0.5$ และ 0.9 การคัดกรองตัวแปรด้วยวิธี Adaptive Lasso จะมีค่ามากที่สุดและเข้าใกล้ $|S|$ มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วยวิธี Adaptive Lasso จึงเหมาะสมมากที่สุด
- เมื่อ $|S|$ มี 50 ค่า และตัวแปรอิสระมีค่าความสัมพันธ์ที่ $\rho = 0$ และ 0.9 การคัดกรองตัวแปรด้วยวิธี SCAD สามารถหาค่า $|\hat{S}|$ ได้เท่ากันซึ่งมีค่ามากที่สุดและเข้าใกล้ $|S|$ มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วยวิธี SCAD จึงมีความเหมาะสมมากที่สุด และที่ $\rho = 0.5$ การคัดกรองตัวแปรด้วยวิธี Adaptive Lasso และ SCAD จะมีค่ามากที่สุดและเข้าใกล้ $|S|$ มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วยวิธี Adaptive Lasso และ SCAD จึงเหมาะสมมากที่สุด

2.) ที่จำนวนตัวแปรอิสระ (p) เท่ากับ 500

- เมื่อ $|S|$ มี 10 ค่า และตัวแปรอิสระมีค่าความสัมพันธ์ที่ $\rho = 0, 0.5$ และ 0.9 การคัดกรองตัวแปรด้วยวิธี วิธี SCAD สามารถหาค่า $|\hat{S}|$ ได้มากที่สุดและเข้าใกล้ $|S|$ มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วยวิธี วิธี SCAD จึงมีความเหมาะสมมากที่สุด
- เมื่อ $|S|$ มี 20 ค่า และตัวแปรอิสระมีค่าความสัมพันธ์ที่ $\rho = 0, 0.5$ และ 0.9 การคัดกรองตัวแปรด้วยวิธี วิธี Adaptive Lasso สามารถหาค่า $|\hat{S}|$ ได้มากที่สุดและเข้าใกล้ $|S|$ มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วยวิธี วิธี Adaptive Lasso จึงมีความเหมาะสมมากที่สุด
- เมื่อ $|S|$ มี 50 ค่า และตัวแปรอิสระมีค่าความสัมพันธ์ที่ $\rho = 0$ การคัดกรองตัวแปรด้วยวิธี Adaptive Lasso และ SCAD สามารถหาค่า $|\hat{S}|$ ได้มากที่สุดและเข้าใกล้ $|S|$ มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วยวิธี Adaptive Lasso และ SCAD จึงมีความเหมาะสมมากที่สุด และที่ $\rho = 0.5$ การคัดกรองตัวแปรด้วยวิธี Adaptive Lasso จะมีค่ามากที่สุดและ

เข้าใกล้ $|S|$ มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วยวิธี Adaptive Lasso จึงเหมาะสมมากที่สุด และที่ $\rho = 0.9$ การคัดกรองตัวแปรด้วยวิธี SCAD จะมีค่ามากที่สุดและเข้าใกล้ $|S|$ มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วยวิธี SCAD จึงเหมาะสมมากที่สุด

3.) ที่จำนวนตัวแปรอิสระ (p) เท่ากับ 1000

- เมื่อ $|S|$ มี 10 ค่า และตัวแปรอิสระมีค่าความสัมพันธ์ที่ $\rho = 0, 0.5$ และ 0.9 การคัดกรองตัวแปรด้วยวิธี วิธี SCAD สามารถหาค่า $|\hat{S}|$ ได้มากที่สุดและเข้าใกล้ $|S|$ มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วยวิธี วิธี SCAD จึงมีความเหมาะสมมากที่สุด
- เมื่อ $|S|$ มี 20 ค่า และตัวแปรอิสระมีค่าความสัมพันธ์ที่ $\rho = 0, 0.5$ และ 0.9 การคัดกรองตัวแปรด้วยวิธี วิธี Adaptive Lasso สามารถหาค่า $|\hat{S}|$ ได้มากที่สุดและเข้าใกล้ $|S|$ มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วยวิธี วิธี Adaptive Lasso จึงมีความเหมาะสมมากที่สุด
- เมื่อ $|S|$ มี 20 ค่า และตัวแปรอิสระมีค่าความสัมพันธ์ที่ $\rho = 0$ ไม่มีวิธีการคัดกรองตัวแปรใดที่เหมาะสมในกรณีนี้ และที่ $\rho = 0.5$ และ 0.9 การคัดกรองตัวแปรด้วยวิธี วิธี Adaptive Lasso สามารถหาค่า $|\hat{S}|$ ได้มากที่สุดและเข้าใกล้ $|S|$ มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วยวิธี วิธี Adaptive Lasso จึงมีความเหมาะสมมากที่สุด

และจากผลในตารางที่ 4.2.2 ยังสามารถสรุปได้อีกว่า

- ในกรณีที่ p มีจำนวนเพิ่มขึ้นจะทำให้ประสิทธิภาพในการหาค่า $|\hat{S}|$ ให้ใกล้เคียงกับ $|S|$ ลดลง
- ในกรณีที่ ρ มีค่าเพิ่มขึ้นจะทำให้ประสิทธิภาพในการหาค่า $|\hat{S}|$ ให้ใกล้เคียงกับ $|S|$ ลดลง
- ในกรณีที่ $|S|$ มีจำนวนเพิ่มขึ้นจะทำให้ประสิทธิภาพในการหาค่า $|\hat{S}|$ ให้ใกล้เคียงกับ $|S|$ ลดลง

และจากตารางที่ 4.2.1 และตารางที่ 4.2.2 จะได้ว่าขนาดของสัมประสิทธิ์ที่ไม่เท่ากับ 0 ที่ขนาดใหญ่ (Large effect Size) จะมีประสิทธิภาพในการหาค่า $|\hat{S}|$ ให้ใกล้เคียงกับ $|S|$ ได้ดีกว่าขนาดเล็ก (Small effect Size)

ตารางที่ 4.2.3 แสดงค่าเฉลี่ย(ค่าเบี่ยงเบนมาตรฐาน) ของ FP เมื่อควบคุม FDR ที่ระดับ 0.1 โดยคำนวณจากข้อมูล 100 ชุด กรณีที่ขนาดตัวอย่าง (n) เท่ากับ 100 และขนาดของสัมประสิทธิ์ที่ไม่เท่ากับ 0 มีขนาดเล็ก (Small effect Size)

n : p	|S|	ขนาด(Effect size) ของสัมประสิทธิ์ที่ไม่เท่ากับ 0 มีขนาดเล็ก $(0 < \|\beta\| < 1)$											
		ความสัมพันธ์ (Correlation) ของตัวแปรอิสระ											
		$\rho = 0$				$\rho = 0.5$				$\rho = 0.9$			
		Lasso	Adaptive Lasso	EN	SCAD	Lasso	Adaptive Lasso	EN	SCAD	Lasso	Adaptive Lasso	EN	SCAD
100:200	10	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0.010 (0.100)	0 (0.000)	0.020 (0.141)	0 (0.000)	0.560 (0.808)	0 (0.000)	0.770 (1.118)
	20	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0.500 (0.835)	0 (0.000)	0.720 (0.899)
	50	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0.070 (0.256)	0 (0.000)	0.390 (0.665)
100:500	10	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0.020 (0.141)	0.570 (0.807)	0 (0.000)	0.390 (0.737)
	20	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0.260 (0.543)	0 (0.000)	0,220 (0.542)
	50	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0.070 (0.293)	0 (0.000)	0.060 (0.0278)
100:1000	10	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0.030 (0.171)	0.500 (0.905)	0 (0.000)	0.190 (0.464)
	20	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0.130 (0.485)	0 (0.000)	0.060 (0.238)
	50	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0.080 (0.307)	0 (0.000)	0.050 (0.219)

หมายเหตุ ช่องที่ระบายสี หมายถึง วิธีที่เหมาะสมที่สุดในแต่ละกรณี

จากตารางที่ 4.2.3 ซึ่งแสดงผลของ False Positive โดยเฉลี่ยของข้อมูลจำลองขนาด 100 ระหว่างการคัดกรองตัวแปรด้วยวิธี Lasso, Adaptive Lasso, En และ SCAD โดยเปรียบเทียบกับจำนวนสัมประสิทธิ์ของตัวแบบที่แท้จริง ที่ขนาดของสัมประสิทธิ์ไม่เท่ากับ 0 มีขนาดเล็ก $(0 < |\beta| < 1)$ พบว่า

1.) ที่จำนวนตัวแปรอิสระ (p) เท่ากับ 200
   - เมื่อ $|S|$ มี 10 ค่า และตัวแปรอิสระมีค่าความสัมพันธ์ที่ $\rho = 0$ การคัดกรองตัวแปรด้วยทั้ง 4 วิธีข้างต้นให้ค่า FP เท่ากับ 0 ทั้งหมด และที่ $\rho = 0.5$ และ 0.9 การคัดกรองตัวแปรด้วยวิธี Lasso และ EN จะมีค่า FP ต่ำที่สุดและใกล้ 0 มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วยวิธี Lasso และ EN จึงเหมาะสมมากที่สุด
   - เมื่อ $|S|$ มี 20 และ 50 ค่า และตัวแปรอิสระมีค่าความสัมพันธ์ที่ $\rho = 0$ และ 0.5 การคัดกรองตัวแปรด้วยทั้ง 4 วิธีข้างต้นให้ค่า FP เท่ากับ 0 ทั้งหมด และที่ $\rho = 0.9$ การคัดกรองตัวแปรด้วยวิธี Lasso และ EP จะมีค่า FP ต่ำที่สุดและใกล้ 0 มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วยวิธี Lasso และ EN จึงเหมาะสมมากที่สุด

2.) ที่จำนวนตัวแปรอิสระ (p) เท่ากับ 500
   - เมื่อ $|S|$ มี 10 ค่า และตัวแปรอิสระมีค่าความสัมพันธ์ที่ $\rho = 0$ และ 0.5 การคัดกรองตัวแปรด้วยทั้ง 4 วิธีข้างต้นให้ค่า FP เท่ากับ 0 ทั้งหมด และที่ $\rho = 0.9$ การคัดกรองตัวแปรด้วยวิธี EN จะมีค่า FP ต่ำที่สุดและใกล้ 0 มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วยวิธี EN จึงเหมาะสมมากที่สุด
   - เมื่อ $|S|$ มี 20 และ 50 ค่า และตัวแปรอิสระมีค่าความสัมพันธ์ที่ $\rho = 0$ และ 0.5 การคัดกรองตัวแปรด้วยทั้ง 4 วิธีข้างต้นให้ค่า FP เท่ากับ 0 ทั้งหมด และที่ $\rho = 0.9$ การคัดกรองตัวแปรด้วยวิธี Lasso และ EN จะมีค่า FP ต่ำที่สุดและใกล้ 0 มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วยวิธี Lasso และ EN จึงเหมาะสมมากที่สุด

3.) ที่จำนวนตัวแปรอิสระ (p) เท่ากับ 1000
   - เมื่อ $|S|$ มี 10 ค่า และตัวแปรอิสระมีค่าความสัมพันธ์ที่ $\rho = 0$ และ 0.5 การคัดกรองตัวแปรด้วยทั้ง 4 วิธีข้างต้นให้ค่า FP เท่ากับ 0 ทั้งหมด และที่ $\rho = 0.9$ การคัดกรองตัวแปรด้วยวิธี EN จะมีค่า FP ต่ำที่สุดและใกล้ 0 มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วยวิธี EN จึงเหมาะสมมากที่สุด
   - เมื่อ $|S|$ มี 20 และ 50 ค่า และตัวแปรอิสระมีค่าความสัมพันธ์ที่ $\rho = 0$ และ 0.5 การคัดกรองตัวแปรด้วยทั้ง 4 วิธีข้างต้นให้ค่า FP เท่ากับ 0 ทั้งหมด และที่ $\rho = 0.9$ การคัดกรองตัวแปรด้วยวิธี Lasso และ EN จะมีค่า FP ต่ำที่สุดและใกล้ 0 มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วยวิธี Lasso และ EN จึงเหมาะสมมากที่สุด

และจากผลในตารางที่ 4.2.3 ยังสามารถสรุปได้อีกว่า

- ในกรณีที่ p มีจำนวนเพิ่มขึ้นจะทำให้ประสิทธิภาพในการหาค่า FP ให้เข้าใกล้ 0 เพิ่มขึ้น
- ในกรณีที่ ρ มีค่าเพิ่มขึ้นจะทำให้ประสิทธิภาพในการหาค่า FP ให้เข้าใกล้ 0 ลดลง
- ในกรณีที่ $|S|$ มีจำนวนเพิ่มขึ้นจะทำให้ประสิทธิภาพในการหาค่า FP ให้เข้าใกล้ 0 เพิ่มขึ้น

ตารางที่ 4.2.4 แสดงค่าเฉลี่ย(ค่าเบี่ยงเบนมาตรฐาน) ของ FP เมื่อควบคุม FDR ที่ระดับ 0.1 โดยคำนวณจากข้อมูล 100 ชุด กรณีที่ขนาดตัวอย่าง (n) เท่ากับ 100 และขนาดของสัมประสิทธิ์ที่ไม่เท่ากับ 0 มีขนาดใหญ่ (Large effect Size)

n : p	|S|	ขนาด(Effect size) ของสัมประสิทธิ์ที่ไม่เท่ากับ 0 มีขนาดใหญ่ $(1 < \|\beta\| < 10)$											
		ความสัมพันธ์ (Correlation) ของตัวแปรอิสระ											
		$\rho = 0$				$\rho = 0.5$				$\rho = 0.9$			
		Lasso	Adaptive Lasso	EN	SCAD	Lasso	Adaptive Lasso	EN	SCAD	Lasso	Adaptive Lasso	EN	SCAD
100:200	10	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0.010 (0.100)	0 (0.000)	0 (0.000)	0 (0.000)	1.050 (1.226)	0 (0.000)	3.810 (2.097)
	20	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0.010 (0.100)	0 (0.000)	0.420 (0.638)	0 (0.000)	2.400 (1.907)
	50	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0.040 (0.243)	0 (0.000)	0.620 (1.002)
100:500	10	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0.010 (0.100)	2.090 (1.589)	0 (0.000)	3.850 (1.866)
	20	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0.340 (0.713)	0 (0.000)	0.410 (0.712)
	50	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0.020 (0.140)	0 (0.000)	0.090 (0.320)
100:1000	10	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0.020 (0.141)	2.650 (1.641)	0 (0.000)	3.010 (1.654)
	20	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0.160 (0.443)	0 (0.000)	0.150 (0.479)
	50	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0.010 (0.100)	0.020 (0.141)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)

หมายเหตุ ช่องที่ระบายสี หมายถึง วิธีที่เหมาะสมที่สุดในแต่ละกรณี

จากตารางที่ 4.2.4 ซึ่งแสดงผลของ False Positive โดยเฉลี่ยของข้อมูลจำลองขนาด 100 ระหว่างการคัดกรองตัวแปรด้วยวิธี Lasso, Adaptive Lasso, En และ SCAD โดยเปรียบเทียบกับจำนวนสัมประสิทธิ์ของตัวแบบที่แท้จริง ที่ขนาดของสัมประสิทธิ์ไม่เท่ากับ 0 มีขนาดใหญ่ $(1 < |\beta| < 10)$ พบว่า

1.) ที่จำนวนตัวแปรอิสระ (p) เท่ากับ 200

- เมื่อ $|S|$ มี 10 ค่า และตัวแปรอิสระมีค่าความสัมพันธ์ที่ $\rho = 0$ การคัดกรองตัวแปรด้วยทั้ง 4 วิธีข้างต้นให้ค่า FP เท่ากับ 0 ทั้งหมด ที่ $\rho = 0.5$ การคัดกรองตัวแปรด้วยวิธี Lasso, EN และ SCAD จะมีค่า FP ต่ำที่สุดและใกล้ 0 มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วย Lasso, EN และ SCAD จึงเหมาะสมที่สุด และที่ $\rho = 0.9$ การคัดกรองตัวแปรด้วยวิธี Lasso และ EN จะมีค่า FP ต่ำที่สุดและใกล้ 0 มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วย Lasso และ EN จึงเหมาะสมที่สุด
- เมื่อ $|S|$ มี 20 ค่า และตัวแปรอิสระมีค่าความสัมพันธ์ที่ $\rho = 0$ การคัดกรองตัวแปรด้วยทั้ง 4 วิธีข้างต้นให้ค่า FP เท่ากับ 0 ทั้งหมด ที่ $\rho = 0.5$ การคัดกรองตัวแปรด้วยวิธี Lasso, Adaptive Lasso และ EN จะมีค่า FP ต่ำที่สุดและใกล้ 0 มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วย Lasso, Adaptive Lasso และ EN จึงเหมาะสมที่สุด และที่ $\rho = 0.9$ การคัดกรองตัวแปรด้วยวิธี Lasso และ EN จะมีค่า FP ต่ำที่สุดและใกล้ 0 มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วย Lasso และ EN จึงเหมาะสมที่สุด
- เมื่อ $|S|$ มี 50 ค่า และตัวแปรอิสระมีค่าความสัมพันธ์ที่ $\rho = 0$ และ 0.5 การคัดกรองตัวแปรด้วยทั้ง 4 วิธีข้างต้นให้ค่า FP เท่ากับ 0 ทั้งหมด และที่ $\rho = 0.9$ การคัดกรองตัวแปรด้วยวิธี Lasso และ EP จะมีค่า FP ต่ำที่สุดและใกล้ 0 มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วยวิธี Lasso และ EN จึงเหมาะสมที่สุด

2.) ที่จำนวนตัวแปรอิสระ (p) เท่ากับ 500

- เมื่อ $|S|$ มี 10 ค่า และตัวแปรอิสระมีค่าความสัมพันธ์ที่ $\rho = 0$ และ 0.5 การคัดกรองตัวแปรด้วยทั้ง 4 วิธีข้างต้นให้ค่า FP เท่ากับ 0 ทั้งหมด และที่ $\rho = 0.9$ การคัดกรองตัวแปรด้วยวิธี EN จะมีค่า FP ต่ำที่สุดและใกล้ 0 มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วยวิธี EN จึงเหมาะสมที่สุด
- เมื่อ $|S|$ มี 20 และ 50 ค่า และตัวแปรอิสระมีค่าความสัมพันธ์ที่ $\rho = 0$ และ 0.5 การคัดกรองตัวแปรด้วยทั้ง 4 วิธีข้างต้นให้ค่า FP เท่ากับ 0 ทั้งหมด และที่ $\rho = 0.9$ การคัดกรองตัวแปรด้วยวิธี Lasso และ EN จะมีค่า FP ต่ำที่สุดและใกล้ 0 มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วยวิธี Lasso และ EN จึงเหมาะสมที่สุด

3.) ที่จำนวนตัวแปรอิสระ (p) เท่ากับ 1000

- เมื่อ $|S|$ มี 10 ค่า และตัวแปรอิสระมีค่าความสัมพันธ์ที่ $\rho = 0$ และ 0.5 การคัดกรองตัวแปรด้วยทั้ง 4 วิธีข้างต้นให้ค่า FP เท่ากับ 0 ทั้งหมด และที่ $\rho = 0.9$ การคัดกรองตัวแปรด้วยวิธี

EN จะมีค่า FP ต่ำที่สุดและใกล้ 0 มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วยวิธี EN จึงเหมาะสมที่สุด

- เมื่อ $|S|$ มี 20 ค่า และตัวแปรอิสระมีค่าความสัมพันธ์ที่ $\rho = 0$ และ 0.5 การคัดกรองตัวแปรด้วยทั้ง 4 วิธีข้างต้นให้ค่า FP เท่ากับ 0 ทั้งหมด ที่ $\rho = 0.9$ การคัดกรองตัวแปรด้วยวิธี Lasso และ EN จะมีค่า FP ต่ำที่สุดและใกล้ 0 มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วยวิธี Lasso และ EN จึงเหมาะสมที่สุด
- เมื่อ $|S|$ มี 50 ค่า และตัวแปรอิสระมีค่าความสัมพันธ์ที่ $\rho = 0$ และ 0.9 การคัดกรองตัวแปรด้วยทั้ง 4 วิธีข้างต้นให้ค่า FP เท่ากับ 0 ทั้งหมด ที่ $\rho = 0.5$ การคัดกรองตัวแปรด้วยวิธี EN และ SCAD จะมีค่า FP ต่ำที่สุดและใกล้ 0 มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วยวิธี EN และ SCAD จึงเหมาะสมที่สุด

และจากผลในตารางที่ 4.2.4 ยังสามารถสรุปได้อีกว่า

- ในกรณีที่ p มีจำนวนเพิ่มขึ้นจะทำให้ประสิทธิภาพในการหาค่า FP ให้เข้าใกล้ 0 จะลดลง
- ในกรณีที่ ρ มีค่าเพิ่มขึ้นจะทำให้ประสิทธิภาพในการหาค่า FP ให้เข้าใกล้ 0 จะลดลง
- ในกรณีที่ $|S|$ มีจำนวนเพิ่มขึ้นจะทำให้ประสิทธิภาพในการหาค่า FP ให้เข้าใกล้ 0 จะเพิ่มขึ้น

และจากตารางที่ 4.2.3 และตารางที่ 4.2.4 จะได้ว่าขนาดของสัมประสิทธิ์ที่ไม่เท่ากับ 0 ที่ขนาดใหญ่ (Large effect Size) จะมีประสิทธิภาพในการหาค่า FN ให้เข้าใกล้ 0 น้อยกว่าขนาดเล็ก (Small effect Size)

ตารางที่ 4.2.5 แสดงค่าเฉลี่ย(ค่าเบี่ยงเบนมาตรฐาน) ของ FN เมื่อควบคุม FDR ที่ระดับ 0.1 โดยคำนวณจากข้อมูล 100 ชุด กรณีที่ขนาดตัวอย่าง (n) เท่ากับ 100 และขนาดของสัมประสิทธิ์ที่ไม่เท่ากับ 0 มีขนาดเล็ก (Small effect Size)

n : p	|S|	ขนาด(Effect size) ของสัมประสิทธิ์ที่ไม่เท่ากับ 0 มีขนาดเล็ก $(0 < \|\beta\| < 1)$											
		ความสัมพันธ์ (Correlation) ของตัวแปรอิสระ											
		$\rho = 0$				$\rho = 0.5$				$\rho = 0.9$			
		Lasso	Adaptive Lasso	EN	SCAD	Lasso	Adaptive Lasso	EN	SCAD	Lasso	Adaptive Lasso	EN	SCAD
100:200	10	8.090	7.130	9.810	7.120	8.310	7.080	9.790	7.140	9.750	8.060	10.000	8.130
		(1.443)	(1.612)	(0.464)	(1.545)	(1.383)	(1.667)	(0.573)	(1.669)	(0.500)	(1.496)	(0.000)	(1.467)
	20	19.700	18.850	19.990	19.040	19.770	19.000	19.990	19.180	19.960	18.420	20.000	18.390
		(0.703)	(1.765)	(0.100)	(1.325)	(0.679)	(1.524)	(0.100)	(1.250)	(0.196)	(1.512)	(0.000)	(1.588)
	50	50,000	49.990	50,000	49.990	50,000	49.930	50,000	50,000	49.990	49.210	50,000	48.570
		(0.000)	(0.100)	(0.000)	(0.100)	(0.000)	(0.256)	(0.000)	(0.000)	(0.100)	(1.075)	(0.000)	(1.748)
100:500	10	10.000	8.070	9.920	8.160	8.610	7.980	9.940	8.070	9.710	8.350	9.990	8.750
		(0.000)	(1.471)	(0.307)	(1.419)	(1.238)	(1.705)	(0.238)	(1.590)	(0.686)	(1.465)	(0.100)	(1.217)
	20	19.840	19.710	19.980	19.700	19.880	19.760	20.000	19.700	19.940	19.330	20.000	19.490
		(0.465)	(0.640)	(0.140)	(0.611)	(0.432)	(0.683)	(0.000)	(0.703)	(0.238)	(1.025)	(0.000)	(0.810)
	50	49.980	49.980	50.000	49.970	49.990	49.970	50,000	50,000	50,000	49.890	50,000	49.930
		(0.141)	(0.141)	(0.000)	(0.171)	(0.100)	(0.222)	(0.000)	(0.000)	(0.000)	(0.345)	(0.000)	(0.293)
100:1000	10	9.090	8.570	9.960	8.680	9.140	8.760	9.970	8.720	9.700	8.650	10.000	9.040
		(1.120)	(1.416)	(0.242)	(1.391)	(1.082)	(1.371)	(0.171)	(1.333)	(0.627)	(1.388)	(0.000)	(1.179)
	20	19.900	19.730	19.990	19.740	19.950	19.810	20.000	19.870	19.970	19.720	20.000	19.880
		(0.362)	(0.708)	(0.100)	(0.645)	(0.219)	(0.563)	(0.000)	(0.393)	(0.171)	(0.604)	(0.000)	(0.356)
	50	50,000	50,000	50,000	50,000	49.990	49.990	50,000	49.990	50,000	49.910	50,000	49.950
		(0.000)	(0.000)	(0.000)	(0.000)	(0.100)	(0.100)	(0.000)	(0.100)	(0.000)	(0.320)	(0.000)	(0.261)

หมายเหตุ ช่องที่ระบายสี หมายถึง วิธีที่เหมาะสมที่สุดในแต่ละกรณี

จากตารางที่ 4.2.5 ซึ่งแสดงผลของ False Negative โดยเฉลี่ยของข้อมูลจำลองขนาด 100 ระหว่างการคัดกรองตัวแปรด้วยวิธี Lasso, Adaptive Lasso, En และ SCAD โดยเปรียบเทียบกับจำนวนสัมประสิทธิ์ของตัวแบบที่แท้จริง ที่ขนาดของสัมประสิทธิ์ไม่เท่ากับ 0 มีขนาดเล็ก $(0 < |\beta| < 1)$ พบว่า

1.) ที่จำนวนตัวแปรอิสระ (p) เท่ากับ 200

- เมื่อ $|S|$ มี 10 ค่า และตัวแปรอิสระมีค่าความสัมพันธ์ที่ $\rho = 0$ การคัดกรองตัวแปรด้วยวิธี SCAD จะมีค่า FN ต่ำที่สุดและเข้าใกล้ 0 มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วย SCAD จึงเหมาะสมที่สุด และที่ $\rho = 0.5$ และ 0.9 การคัดกรองตัวแปรด้วยวิธี Adaptive Lasso จะมีค่า FN ต่ำที่สุดและเข้าใกล้ 0 มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วยวิธี Adaptive Lasso จึงเหมาะสมที่สุด
- เมื่อ $|S|$ มี 20 ค่า และตัวแปรอิสระมีค่าความสัมพันธ์ที่ $\rho = 0$ และ 0.5 การคัดกรองตัวแปรด้วยวิธี Adaptive Lasso จะมีค่า FN ต่ำที่สุดและเข้าใกล้ 0 มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วยวิธี Adaptive Lasso จึงเหมาะสมที่สุด และที่ $\rho = 0.9$ การคัดกรองตัวแปรด้วยวิธี SCAD จะมีค่า FN ต่ำที่สุดและเข้าใกล้ 0 มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วย SCAD จึงเหมาะสมที่สุด
- เมื่อ $|S|$ มี 50 ค่า และตัวแปรอิสระมีค่าความสัมพันธ์ที่ $\rho = 0$ การคัดกรองตัวแปรด้วยวิธี Adaptive Lasso และ SCAD จะมีค่า FN ต่ำที่สุดและเข้าใกล้ 0 มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วยวิธี Adaptive Lasso และ SCAD จึงเหมาะสมที่สุด ที่ $\rho = 0.5$ การคัดกรองตัวแปรด้วยวิธี Adaptive Lasso จะมีค่า FN ต่ำที่สุดและเข้าใกล้ *0* มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วยวิธี Adaptive Lasso จึงเหมาะสมที่สุด และที่ $\rho = 0.9$ การคัดกรองตัวแปรด้วยวิธี SCAD จะมีค่า FN ต่ำที่สุดและเข้าใกล้ 0 มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วยวิธี SCAD จึงเหมาะสมที่สุด

2.) ที่จำนวนตัวแปรอิสระ (p) เท่ากับ 500

- เมื่อ $|S|$ มี 10 ค่า และตัวแปรอิสระมีค่าความสัมพันธ์ที่ $\rho = 0$, 0.5 และ 0.9 การคัดกรองตัวแปรด้วยวิธี Adaptive Lasso จะมีค่า FN ต่ำที่สุดและเข้าใกล้ 0 มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วย Adaptive Lasso จึงเหมาะสมที่สุด
- เมื่อ $|S|$ มี 20 ค่า และตัวแปรอิสระมีค่าความสัมพันธ์ที่ $\rho = 0$ และ 0.5 การคัดกรองตัวแปรด้วยวิธี SCAD จะมีค่า FN ต่ำที่สุดและเข้าใกล้ 0 มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วยวิธี SCAD จึงเหมาะสมที่สุด และที่ $\rho = 0.9$ การคัดกรองตัวแปรด้วยวิธี Adaptive Lasso จะมีค่า FN ต่ำที่สุดและเข้าใกล้ 0 มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วย Adaptive Lasso จึงเหมาะสมที่สุด
- เมื่อ $|S|$ มี 50 ค่า และตัวแปรอิสระมีค่าความสัมพันธ์ที่ $\rho = 0$ การคัดกรองตัวแปรด้วยวิธี SCAD จะมีค่า FN ต่ำที่สุดและเข้าใกล้ 0 มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วยวิธี

SCAD จึงเหมาะสมที่สุด และที่ ρ =0.5 และ 0.9 การคัดกรองตัวแปรด้วยวิธี Adaptive Lasso จะมีค่า FN ต่ำที่สุดและเข้าใกล้ 0 มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วย Adaptive Lasso จึงเหมาะสมที่สุด

3.) ที่จำนวนตัวแปรอิสระ (p) เท่ากับ 1000

- เมื่อ $|S|$ มี 10 ค่า และตัวแปรอิสระมีค่าความสัมพันธ์ที่ ρ = 0 และ 0.9 การคัดกรองตัวแปรด้วยวิธี Adaptive Lasso จะมีค่า FN ต่ำที่สุดและเข้าใกล้ 0 มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วยวิธี Adaptive Lasso จึงเหมาะสมที่สุด และที่ ρ = 0.5 การคัดกรองตัวแปรด้วยวิธี SCAD จะมีค่า FN ต่ำที่สุดและเข้าใกล้ 0 มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วย SCAD จึงเหมาะสมที่สุด
- เมื่อ $|S|$ มี 20 ค่า และตัวแปรอิสระมีค่าความสัมพันธ์ที่ ρ = 0, 0.5 และ 0.9 การคัดกรองตัวแปรด้วยวิธี Adaptive Lasso จะมีค่า FN ต่ำที่สุดและเข้าใกล้ 0 มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วย Adaptive Lasso จึงเหมาะสมที่สุด
- เมื่อ $|S|$ มี 50 ค่า และตัวแปรอิสระมีค่าความสัมพันธ์ที่ ρ = 0 การคัดกรองตัวแปรทั้ง 4 วิธีข้างต้นให้ค่า FN เท่ากันทั้งหมด ที่ ρ =0.5 การคัดกรองตัวแปรด้วยวิธี Lasso, Adaptive Lasso และ SCAD จะมีค่า FN ต่ำที่สุดและเข้าใกล้ 0 มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วย Lasso, Adaptive Lasso และ SCAD จึงเหมาะสมที่สุด และที่ ρ =0.9 การคัดกรองตัวแปรด้วยวิธี Adaptive Lasso จะมีค่า FN ต่ำที่สุดและเข้าใกล้ 0 มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรวิธี Adaptive Lasso จึงเหมาะสมที่สุด

และจากผลในตารางที่ 4.2.5 ยังสามารถสรุปได้อีกว่า

- ในกรณีที่ p มีจำนวนเพิ่มขึ้นจะทำให้ประสิทธิภาพในการหาค่า FN ให้เข้าใกล้ 0 ลดลง
- ในกรณีที่ ρ มีค่าเพิ่มขึ้นจะทำให้ประสิทธิภาพในการหาค่า FN ให้เข้าใกล้ 0 ลดลง
- ในกรณีที่ $|S|$ มีจำนวนเพิ่มขึ้นจะทำให้ประสิทธิภาพในการหาค่า FN ให้เข้าใกล้ 0 ลดลง

ตารางที่ 4.2.6 แสดงค่าเฉลี่ย(ค่าเบี่ยงเบนมาตรฐาน) ของ FP เมื่อควบคุม FDR ที่ระดับ 0.1 โดยคำนวณจากข้อมูล 100 ชุด กรณีที่ขนาดตัวอย่าง (n) เท่ากับ 100 และขนาดของสัมประสิทธิ์ที่ไม่เท่ากับ 0 มีขนาดใหญ่ (Large effect Size)

n : p	|S|	ขนาด(Effect size) ของสัมประสิทธิ์ที่ไม่เท่ากับ 0 มีขนาดใหญ่ $(1 < \|\beta\| < 10)$											
		ความสัมพันธ์ (Correlation) ของตัวแปรอิสระ											
		$\rho = 0$				$\rho = 0.5$				$\rho = 0.9$			
		Lasso	Adaptive Lasso	EN	SCAD	Lasso	Adaptive Lasso	EN	SCAD	Lasso	Adaptive Lasso	EN	SCAD
100:200	10	0.140	0.000	2.440	0.000	0.360	0.000	3.610	0.000	1.980	0.430	6.270	0.710
		(0.376)	(0.000)	(2.257)	(0.000)	(0.811)	(0.000)	(2.915)	(0.000)	(1.428)	(0.831)	(3.428)	(1.056)
	20	16.950	11.950	19.720	15.960	17.140	11.570	19.980	15.520	19.220	10.510	20.000	13.660
		(4.205)	(5.390)	(1.247)	(4.494)	(3.920)	(5.039)	(0.140)	(4.491)	(1.850)	(3.721)	(0.000)	(2.832)
	50	49.950	49.940	50.000	49.920	50.000	49.990	50.000	49.990	50.000	49.650	50.000	48.200
		(0.261)	(0.238)	(0.000)	(0.272)	(0.000)	(0.100)	(0.000)	(0.100)	(0.000)	(0.891)	(0.000)	(1.964)
100:500	10	0.860	0.410	7.600	0.000	1.740	0.820	8.500	0.010	3.850	0.970	9.920	0.900
		(1.428)	(0.817)	(2.403)	(0.000)	(2.111)	(1.297)	(2.492)	(0.100)	(2.847)	(1.298)	(0.706)	(1.210)
	20	19.820	18.550	19.990	19.090	19.710	18.540	20/000	19.070	19/960	16.760	20.000	18.280
		(0.557)	(2.328)	(0.100)	(1.491)	(0.820)	(2.199)	(0.000)	(1.545)	(0.196)	(3.062)	(0.000)	(1.990)
	50	49.990	49.980	50.000	49/980	49.990	49.980	50.000	49.980	50.000	49.920	50.000	49.850
		(0.100)	(0.141)	(0.000)	(0.141)	(0.100)	(0.1410)	(0.000)	(0.200)	(0.000)	(0.338)	(0.000)	(0.457)
100:1000	10	3.570	2.240	9.030	0.260	4.200	2.940	9.340	0.470	5.350	1.780	9.960	1.450
		(2.404)	(1.848)	(1.374)	(1.219)	(2.704)	(2.228)	(1.408)	(1.855)	(3.098)	(1.856)	(0.242)	(2.066)
	20	19.840	19.670	20.000	19.710	19.900	19.530	19.990	19.730	19.940	19.060	20.000	19.440
		(0.373)	(0.804)	(0.000)	(0.868)	(0.333)	(1.258)	(0.100)	(0.633)	(0.277)	(1.536)	(0.000)	(1.094)
	50	50.000	50.000	50.000	50.000	49.990	49.980	50.000	50.000	49.980	49.970	50.000	49.980
		(0.000)	(0.000)	(0.000)	(0.000)	(0.100)	(0.141)	(0.000)	(0.000)	(0.141)	(0.222)	(0.000)	(0.141)

หมายเหตุ ช่องที่ระบายสี หมายถึง วิธีที่เหมาะสมที่สุดในแต่ละกรณี

จากตารางที่ 4.2.6 ซึ่งแสดงผลของ False Negative โดยเฉลี่ยของข้อมูลจำลองขนาด 100 ระหว่างการคัดกรองตัวแปรด้วยวิธี Lasso, Adaptive Lasso, En และ SCAD โดยเปรียบเทียบกับจำนวนสัมประสิทธิ์ของตัวแบบที่แท้จริง ที่ขนาดของสัมประสิทธิ์ไม่เท่ากับ 0 มีขนาดใหญ่ $(1 < |\beta| < 10)$ พบว่า

1.) ที่จำนวนตัวแปรอิสระ (p) เท่ากับ 200

- เมื่อ $|S|$ มี 10 ค่า และตัวแปรอิสระมีค่าความสัมพันธ์ที่ $\rho = 0$ และ 0.5 การคัดกรองตัวแปรด้วยวิธี Adaptive Lasso และ SCAD จะมีค่า FN ต่ำที่สุดและเข้า 0 มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วยวิธี Adaptive Lasso และ SCAD จึงเหมาะสมที่สุด และที่ $\rho = 0.9$ การคัดกรองตัวแปรด้วยวิธี Adaptive Lasso จะมีค่า FN ต่ำที่สุดและเข้าใกล้ 0 มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วยวิธี Adaptive Lasso จึงเหมาะสมที่สุด
- เมื่อ $|S|$ มี 20 ค่า และตัวแปรอิสระมีค่าความสัมพันธ์ที่ $\rho = 0$, 0.5 และ 0.9 การคัดกรองตัวแปรด้วยวิธี Adaptive Lasso จะมีค่า FN ต่ำที่สุดและเข้าใกล้ 0 มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วยวิธี Adaptive Lasso จึงเหมาะสมที่สุด
- เมื่อ $|S|$ มี 50 ค่า และตัวแปรอิสระมีค่าความสัมพันธ์ที่ $\rho = 0$ และ 0.9 การคัดกรองตัวแปรด้วยวิธี SCAD จะมีค่า FN ต่ำที่สุดและเข้าใกล้ 0 มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วยวิธี SCAD จึงเหมาะสมที่สุด ที่ $\rho = 0.5$ การคัดกรองตัวแปรด้วยวิธี Adaptive Lasso และ SCAD จะมีค่า FN ต่ำที่สุดและเข้าใกล้ 0 มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วยวิธี Adaptive Lasso และ SCAD จึงเหมาะสมที่สุด

2.) ที่จำนวนตัวแปรอิสระ (p) เท่ากับ 500

- เมื่อ $|S|$ มี 10 ค่า และตัวแปรอิสระมีค่าความสัมพันธ์ที่ $\rho = 0$, 0.5 และ 0.9 การคัดกรองตัวแปรด้วยวิธี SCAD จะมีค่า FN ต่ำที่สุดและเข้าใกล้ 0 มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วย SCAD จึงเหมาะสมที่สุด
- เมื่อ $|S|$ มี 20 ค่า และตัวแปรอิสระมีค่าความสัมพันธ์ที่ $\rho = 0$, 0.5 และ 0.9 การคัดกรองตัวแปรด้วยวิธี Adaptive Lasso จะมีค่า FN ต่ำที่สุดและเข้าใกล้ 0 มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วยวิธี Adaptive Lasso จึงเหมาะสมที่สุด
- เมื่อ $|S|$ มี 50 ค่า และตัวแปรอิสระมีค่าความสัมพันธ์ที่ $\rho = 0$ การคัดกรองตัวแปรด้วยวิธี Adaptive Lasso และ SCAD จะมีค่า FN ต่ำที่สุดและเข้าใกล้ 0 มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วยวิธี Adaptive Lasso และ SCAD จึงเหมาะสมที่สุด ที่ $\rho = 0.5$ การคัดกรองตัวแปรด้วยวิธี Adaptive Lasso จะมีค่า FN ต่ำที่สุดและเข้าใกล้ 0 มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วยวิธี Adaptive Lasso จึงเหมาะสมที่สุด และที่ $\rho = 0.9$ การคัดกรองตัวแปรด้วยวิธี SCAD จะมีค่า FN ต่ำที่สุดและเข้าใกล้ 0 มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วยวิธี SCAD จึงเหมาะสมที่สุด

3.) ที่จำนวนตัวแปรอิสระ (p) เท่ากับ 1000

- เมื่อ $|S|$ มี 10 ค่า และตัวแปรอิสระมีค่าความสัมพันธ์ที่ $\rho = 0, 0.5$ และ 0.9 การคัดกรองตัวแปรด้วยวิธี SCAD จะมีค่า FN ต่ำที่สุดและเข้าใกล้ 0 มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วย SCAD จึงเหมาะสมที่สุด
- เมื่อ $|S|$ มี 20 ค่า และตัวแปรอิสระมีค่าความสัมพันธ์ที่ $\rho = 0, 0.5$ และ 0.9 การคัดกรองตัวแปรด้วยวิธี Adaptive Lasso จะมีค่า FN ต่ำที่สุดและเข้าใกล้ 0 มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วยวิธี Adaptive Lasso จึงเหมาะสมที่สุด
- เมื่อ $|S|$ มี 50 ค่า และตัวแปรอิสระมีค่าความสัมพันธ์ที่ $\rho = 0$ การคัดกรองตัวแปรด้วย 4 วิธีข้างต้นได้ค่า FN เท่ากันทั้งหมด ที่ $\rho = 0.5$ และ 0.9 การคัดกรองตัวแปรด้วยวิธี Adaptive Lasso จะมีค่า FN ต่ำที่สุดและเข้าใกล้ 0 มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วยวิธี Adaptive Lasso จึงเหมาะสมที่สุด

และจากผลในตารางที่ 4.2.6 ยังสามารถสรุปได้อีกว่า

- ในกรณีที่ p มีจำนวนเพิ่มขึ้นจะทำให้ประสิทธิภาพในการหาค่า FN ให้เข้าใกล้ 0 ลดลง
- ในกรณีที่ ρ มีค่าเพิ่มขึ้นจะทำให้ประสิทธิภาพในการหาค่า FN ให้เข้าใกล้ 0 ลดลง
- ในกรณีที่ $|S|$ มีจำนวนเพิ่มขึ้นจะทำให้ประสิทธิภาพในการหาค่า FN ให้เข้าใกล้ 0 ลดลง

และจากตารางที่ 4.2.5 และตารางที่ 4.2.6 จะได้ว่าขนาดของสัมประสิทธิ์ที่ไม่เท่ากับ 0 ที่ขนาดใหญ่ (Large effect Size) จะมีประสิทธิภาพในการหาค่า FN ให้เข้าใกล้ 0 ได้ดีกว่าขนาดเล็ก (Small effect Size)

จากตารางที่ 4.2.1 – 4.2.6 เมื่อพิจารณาจากค่า $|\hat{S}|$, FP และ FN จะได้ว่าเมื่อขนาดตัวอย่าง (n) เท่ากับ 100 ค่าของ $|\hat{S}|$ และ FN จะไปในทิศทางเดียวกัน นั่นคือการคัดกรองตัวแปรด้วยวิธี Adaptive Lasso และวิธี SCAD จะเหมาะสมที่สุด แต่ค่า FP จะได้วิธี Lasso และวิธี EN ที่เหมาะสม นั่นแสดงให้เห็นว่าวิธี Lasso และวิธี EN มีประสิทธิภาพในการคัดเลือกตัวแปรและอำนาจในการทดสอบน้อยกว่าวิธี Adaptive Lasso และวิธี SCAD

4.3 ผลการเปรียบเทียบค่าความผิดพลาดในการตรวจจับเชิงบวก (FP) ค่าความผิดพลาดในการตรวจจับเชิงลบ (FN) และจำนวนของสัมประสิทธิ์ของตัวแปรอิสระที่มีค่าไม่เท่ากับ 0 จากการทดสอบสมมติฐานโดยเฉลี่ย เมื่อควบคุม FDR ที่ระดับ 0.1 ของข้อมูลจำลองขนาด 200 ระหว่างการคัดกรองตัวแปรด้วยวิธี Lasso, Adaptive Lasso, EN และ SCAD

ในส่วนนี้ผู้วิจัยต้องการศึกษาเปรียบเทียบการคัดกรองตัวแปรจากวิธี Lasso, Adaptive Lasso, EN และ SCAD และเพื่อพิจารณาว่าปัจจัยใดที่ส่งผลต่อประสิทธิภาพการทำงานของวิธีการคัดกรองตัวแปรแต่ละวิธี ภายใต้ปัจจัย ดังต่อไปนี้

1. อัตราส่วนระหว่างขนาดตัวอย่างและจำนวนตัวแปรอิสระ คือ 1:2, 1:5 และ 1:10
2. ร้อยละของจำนวนสัมประสิทธิ์ที่ไม่เท่ากับ 0 $(\beta \neq 0)$ เมื่อเทียบกับขนาดตัวอย่าง คือ ร้อยละ 10, 20 และ 50
3. ขนาด (Effect Size) ของสัมประสิทธิ์ที่ไม่เท่ากับ 0 คือ Small Size $(0 < |\beta| < 1)$ และ Large Size $(1 < |\beta| < 10)$
4. ความสัมพันธ์ (Correlation) ของตัวแปรอิสระ คือ ρ = 0, 0.5, 0.9

โดยแสดงผลในตารางที่ 4.3.1 - 4.3.6 โดยแต่ละตารางมีรายละเอียดดังนี้

เกณฑ์ที่ใช้ในการวัด	ตารางที่	ปัจจัยที่ใช้ในการพิจารณา	วิธีการคัดกรองตัวแปรที่ต้องการเปรียบเทียบ
$\|\hat{S}\|$	4.3.1	• n ; p • จำนวนสัมประสิทธิ์ของโมเดลที่แท้จริงที่มีค่าไม่เท่ากับ 0 ($\beta \neq 0$) • ความสัมพันธ์ (Correlation) ของตัวแปรอิสระ	1. Lasso 2. Adaptive Lasso 3. EN 4. SACD
	4.3.2		
FP	4.3.3		
	4.3.4		
FN	4.3.5		
	4.3.6		

ตารางที่ 4.3.1 แสดงค่าเฉลี่ย(ค่าเบี่ยงเบนมาตรฐาน) ของ $|\hat{S}|$ เมื่อควบคุม FDR ที่ระดับ 0.1 โดยคำนวณจากข้อมูล 100 ชุด กรณีที่ขนาดตัวอย่าง (n) เท่ากับ 200 และขนาดของสัมประสิทธิ์ที่ไม่เท่ากับ 0 มีขนาดเล็ก (Small effect Size)

n : p	$\|S\|$	ขนาด(Effect size) ของสัมประสิทธิ์ที่ไม่เท่ากับ 0 มีขนาดเล็ก $(0 < \|\beta\| < 1)$											
		ความสัมพันธ์ (Correlation) ของตัวแปรอิสระ											
		$\rho = 0$				$\rho = 0.5$				$\rho = 0.9$			
		Lasso	Adaptive Lasso	EN	SCAD	Lasso	Adaptive Lasso	EN	SCAD	Lasso	Adaptive Lasso	EN	SCAD
200:400	20	9.120	11.080	2.140	11.370	8.170	10.310	1.440	10.570	2.270	10.830	0.020	12.890
		(1.793)	(2.355)	(2.160)	(2.359)	(2.122)	(2.381)	(1.827)	(2.383)	(1.728)	(3.130)	(0.140)	(3.323)
	40	2.120	6.440	0	4.040	1.280	6.860	0.010	3.890	0.140	9.170	0	9.850
		(2.811)	(4.304)	(0.000)	(3.314)	(1.968)	(4.278)	(0.100)	(3.004)	(0.512)	(3.861)	(0.000)	(4.379)
	100	0.020	0.040	0	0.050	0	0.070	0	0.090	0	2.177	0	5.000
		(0.140)	(0.196)	(0.000)	(0.219)	(0.000)	(0.256)	(0.000)	(0.287)	(0.000)	(2.395)	(0.000)	(4.039)
200:1000	20	7.830	9.450	0.460	9.800	6.080	8.010	0.160	8.160	2.380	11.290	0	10.690
		(2.482)	(2.615)	(0.903)	(2.522)	(2.634)	(2.989)	(0.486)	(2.740)	(1.756)	(3.491)	(0.000)	(3.454)
	40	0.560	1.160	0	1.020	0.400	1.480	0.010	1.290	0.110	3.990	0	2.970
		(0.782)	(1.454)	(0.000)	(1.197)	(1.029)	(2.148)	(0.100)	(2.046)	(0.423)	(3.641)	(0.000)	(2.921)
	100	0.010	0.010	0	0.010	0.010	0.020	0	0.020	0	0.840	0	0.590
		(0.100)	(0.100)	(0.000)	(0.100)	(0.100)	(0.141)	(0.000)	(0.141)	(0.000)	(1.079)	(0.000)	(0.922)
200:2000	20	5.590	7.000	0.290	7.290	4.450	5.860	0.190	5.860	2.010	9.110	0	7.530
		(2.437)	(2.704)	(0.640)	(2.861)	(2.314)	(2.785)	(0.580)	(2.874)	(1.925)	(4.009)	(0.000)	(3.911)
	40	0.270	0.340	0	0.400	0.220	0.470	0.010	0.420	0.080	1.780	0	0.740
		(0.649)	(0.639)	(0.000)	(0.724)	(0.542)	(1.077)	(0.100)	(0.986)	(0.272)	(2.012)	(0.000)	(1.330)
	100	0	0	0	0	0.030	0.030	0	0.030	0.020	0.080	0	0.040
		(0.000)	(0.000)	(0.000)	(0.000)	(0.171)	(0.171)	(0.000)	(0.171)	(0.141)	(0.338)	(0.000)	(0.196)

หมายเหตุ ช่องที่ระบายสี หมายถึง วิธีที่เหมาะสมที่สุดในแต่ละกรณี

จากตารางที่ 4.3.1 ซึ่งแสดงผลของ $|\hat{S}|$ โดยเฉลี่ยของข้อมูลจำลองขนาด 200 ระหว่างการคัดกรองตัวแปรด้วยวิธี Lasso, Adaptive Lasso, En และ SCAD โดยเปรียบเทียบกับจำนวนสัมประสิทธิ์ของตัวแบบที่แท้จริง ที่ขนาดของสัมประสิทธิ์ไม่เท่ากับ 0 มีขนาดเล็ก $(0 < |\beta| < 1)$ พบว่า

1.) ที่จำนวนตัวแปรอิสระ (p) เท่ากับ 400

- เมื่อ $|S|$ มี 20 ค่า เมื่อตัวแปรอิสระมีค่าความสัมพันธ์ที่ $\rho = 0, 0.5$ และ 0.9 การคัดกรองตัวแปรด้วยวิธี SCAD สามารถหาค่า $|\hat{S}|$ ได้มากที่สุดและเข้าใกล้ $|S|$ มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วยวิธี SCAD จึงมีความเหมาะสมที่สุด
- เมื่อ $|S|$ มี 40 ค่า เมื่อตัวแปรอิสระมีค่าความสัมพันธ์ที่ $\rho = 0$ และ 0.5 การคัดกรองตัวแปรด้วยวิธี Adaptive Lasso สามารถหาค่า $|\hat{S}|$ ได้มากที่สุดและเข้าใกล้ $|S|$ มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วยวิธี Adaptive Lasso จึงมีความเหมาะสมที่สุด และที่ $\rho = 0.9$ การคัดกรองตัวแปรด้วยวิธี SCAD จะมีค่ามากที่สุดและเข้าใกล้ $|S|$ มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วยวิธี SCAD จึงเหมาะสมที่สุด
- เมื่อ $|S|$ มี 100 ค่า เมื่อตัวแปรอิสระมีค่าความสัมพันธ์ที่ $\rho = 0, 0.5$ และ 0.9 การคัดกรองตัวแปรด้วยวิธี SCAD สามารถหาค่า $|\hat{S}|$ ได้เท่ากันซึ่งมีค่ามากที่สุดและเข้าใกล้ $|S|$ มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วยวิธี SCAD จึงมีความเหมาะสมที่สุด

2.) ที่จำนวนตัวแปรอิสระ (p) เท่ากับ 1000

- เมื่อ $|S|$ มี 20 ค่า เมื่อตัวแปรอิสระมีค่าความสัมพันธ์ที่ $\rho = 0, 0.5$ และ 0.9 การคัดกรองตัวแปรด้วยวิธี SCAD สามารถหาค่า $|\hat{S}|$ ได้มากที่สุดและเข้าใกล้ $|S|$ มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วยวิธี SCAD จึงมีความเหมาะสมที่สุด
- เมื่อ $|S|$ มี 40 ค่า เมื่อตัวแปรอิสระมีค่าความสัมพันธ์ที่ $\rho = 0, 0.5$ และ 0.9 การคัดกรองตัวแปรด้วยวิธี Adaptive Lasso สามารถหาค่า $|\hat{S}|$ ได้มากที่สุดและเข้าใกล้ $|S|$ มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วยวิธี Adaptive Lasso จึงมีความเหมาะสมที่สุด
- เมื่อ $|S|$ มี 100 ค่า เมื่อตัวแปรอิสระมีค่าความสัมพันธ์ที่ $\rho = 0$ การคัดกรองตัวแปรด้วยวิธี Lasso, Adaptive Lasso และ SCAD สามารถหาค่า $|\hat{S}|$ ได้มากที่สุดและเข้าใกล้ $|S|$ มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วยวิธี Lasso, Adaptive Lasso และ SCAD จึงมีความเหมาะสมที่สุด ที่ $\rho = 0.5$ การคัดกรองตัวแปรด้วยวิธี Adaptive Lasso และ SCAD สามารถหาค่า $|\hat{S}|$ ได้มากที่สุดและเข้าใกล้ $|S|$ มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วยวิธี Adaptive Lasso และ SCAD จึงมีความเหมาะสมที่สุด และที่ $\rho = 0.9$ การคัดกรองตัวแปรด้วยวิธี Adaptive Lasso สามารถหาค่า $|\hat{S}|$ ได้มากที่สุดและเข้าใกล้ $|S|$ มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วยวิธี Adaptive Lasso จึงมีความเหมาะสมที่สุด

3.) ที่จำนวนตัวแปรอิสระ (p) เท่ากับ 2000

- เมื่อ $|S|$ มี 20 และ 40 ค่า เมื่อตัวแปรอิสระมีค่าความสัมพันธ์ที่ $\rho = 0$ การคัดกรองตัวแปรด้วยวิธี SCAD สามารถหาค่า $|\hat{S}|$ ได้มากที่สุดและเข้าใกล้ $|S|$ มากที่สุด ดังนั้นในกรณีนี้การ

คัดกรองตัวแปรด้วยวิธี SCAD จึงมีความเหมาะสมที่สุด และที่ $\rho = 0.5$ และ 0.9 การคัดกรองตัวแปรด้วยวิธี Adaptive Lasso สามารถหาค่า $|\hat{S}|$ ได้มากที่สุดและเข้าใกล้ $|S|$ มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วยวิธี Adaptive Lassoจึงมีความเหมาะสมที่สุด

- เมื่อ $|S|$ มี 100 ค่า เมื่อตัวแปรอิสระมีค่าความสัมพันธ์ที่ $\rho = 0$ การคัดกรองตัวแปรด้วย 4 วิธีข้างต้นให้ค่า $|\hat{S}|$ เท่ากับ 0 ทั้งหมด นั่นคือในทุกวิธีการคัดกรองตัวแปรไม่สามารถหาค่าสัมประสิทธิ์ใดที่ไม่เท่ากับ 0 จากการทดสอบสมมติฐาน ที่ $\rho = 0.5$ การคัดกรองตัวแปรด้วยวิธี Lasso, Adaptive Lasso และ SCAD สามารถหาค่า $|\hat{S}|$ ได้มากที่สุดและเข้าใกล้ $|S|$ มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วยวิธี Lasso, Adaptive Lasso และ SCAD จึงมีความเหมาะสมที่สุด และที่ $\rho = 0.9$ การคัดกรองตัวแปรด้วยวิธี Adaptive Lasso สามารถหาค่า $|\hat{S}|$ ได้มากที่สุดและเข้าใกล้ $|S|$ มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วยวิธี Adaptive Lasso จึงมีความเหมาะสมที่สุด

และจากผลในตารางที่ 4.3.1 ยังสามารถสรุปได้อีกว่า

- ในกรณีที่ p มีจำนวนเพิ่มขึ้นจะทำให้ประสิทธิภาพในการหาค่า $|\hat{S}|$ ให้เข้าใกล้ $|S|$ ลดลง
- ในกรณีที่ ρ มีค่าเพิ่มขึ้นจะทำให้ประสิทธิภาพในการหาค่า $|\hat{S}|$ ให้เข้าใกล้ $|S|$ ลดลง
- ในกรณีที่ $|S|$ มีจำนวนเพิ่มขึ้นจะทำให้ประสิทธิภาพในการหาค่า $|\hat{S}|$ ให้เข้าใกล้ $|S|$ ลดลง

ตารางที่ 4.3.2 แสดงค่าเฉลี่ย(ค่าเบี่ยงเบนมาตรฐาน) ของ $|\hat{S}|$ เมื่อควบคุม FDR ที่ระดับ 0.1 โดยคำนวณจากข้อมูล 100 ชุด กรณีที่ขนาดตัวอย่าง (n) เท่ากับ 200 และขนาดของสัมประสิทธิ์ที่ไม่เท่ากับ 0 มีขนาดใหญ่ (Large effect Size)

n : p	|S|	ขนาด(Effect size) ของสัมประสิทธิ์ที่ไม่เท่ากับ 0 มีขนาดใหญ่ $(1 < \|\beta\| < 10)$											
		ความสัมพันธ์ (Correlation) ของตัวแปรอิสระ											
		$\rho = 0$				$\rho = 0.5$				$\rho = 0.9$			
		Lasso	Adaptive Lasso	EN	SCAD	Lasso	Adaptive Lasso	EN	SCAD	Lasso	Adaptive Lasso	EN	SCAD
200:400	20	20.000	20.000	19.060	20.000	19.950	20.020	18.570	20.000	18.610	21.980	16.960	27.710
		(0.000)	(0.000)	(1.873)	(0.000)	(0.219)	(0.141)	(2.868)	(0.000)	(1.510)	(1.885)	(3.314)	(3.610)
	40	5.740	12.590	0.140	5.030	5.530	14.320	0	5.780	3.480	22.140	0	18.620
		(6.299)	(7.845)	(0.804)	(5.255)	(6.009)	(7.309)	(0.000)	(4.837)	(4.624)	(6.033)	(0.000)	(6.419)
	100	0.010	0.030	0	0.020	0	0.040	0	0.080	0.020	1.460	0	6.840
		(0.100)	(0.171)	(0.000)	(0.141)	(0.000)	(0.242)	(0.000)	(0.307)	(0.141)	(2.061)	(0.000)	(4.545)
200:1000	20	17.700	17.880	5.560	19.980	17.470	18.080	5.860	20.000	17.730	18.940	6.120	20.680
		(2.713)	(1.838)	(3.924)	(0.141)	(2.341)	(1.727)	(4.722)	(0.000)	(2.436)	(1.653)	(4.137)	(0.986)
	40	0.600	1.920	0	1.540	0.540	1.880	0	1.390	0.530	1.760	0	1.310
		(1.385)	(2.501)	(0.000)	(2.090)	(1.395)	(3.127)	(0.000)	(2.024)	(1.213)	(2.882)	(0.000)	(2.356)
	100	0.010	0.030	0	0.020	0	0.040	0	0.020	0	0.380	0	0.600
		(0.100)	(01710)	(0.000)	(0.141)	(0.000)	(0.196)	(0.000)	(0.141)	(0.000)	(0.632)	(0.000)	(1.137)
200:2000	20	11.930	12.830	2.580	19.390	11.670	13.040	1.170	18.890	11.310	13.480	1.790	18.770
		(3.539)	(3.104)	(2.563)	(2.824)	(3.149)	(3.437)	(1.741)	(3.845)	(3.030)	(3.219)	(2.379)	(3.175)
	40	0.260	0.750	0.010	0.600	0.230	0.710	0	0.500	0.120	2.390	0	1.600
		(0.629)	(1.677)	(0.100)	(1.370)	(0.617)	(1.273)	(0.000)	(1.010)	(0.408)	(3.469)	(0.000)	(2.659)
	100	0	0.010	0	0.010	0	0	0	0	0	0.040	0	0.060
		(0.000)	(0.100)	(0.000)	(0.100)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.196)	(0.000)	(0.238)

หมายเหตุ ช่องที่ระบายสี หมายถึง วิธีที่เหมาะสมที่สุดในแต่ละกรณี

จากตารางที่ 4.3.2 ซึ่งแสดงผลของ $|\hat{S}|$ โดยเฉลี่ยของข้อมูลจำลองขนาด 200 ระหว่างการคัดกรองตัวแปรด้วยวิธี Lasso, Adaptive Lasso, En และ SCAD โดยเปรียบเทียบกับจำนวนสัมประสิทธิ์ของตัวแบบที่แท้จริง ที่ขนาดของสัมประสิทธิ์ไม่เท่ากับ 0 มีขนาดใหญ่ $(1 < |\beta| < 10)$ พบว่า

1.) ที่จำนวนตัวแปรอิสระ (p) เท่ากับ 400

- เมื่อ $|S|$ มี 20 ค่า เมื่อตัวแปรอิสระมีค่าความสัมพันธ์ที่ $\rho = 0$ การคัดกรองตัวแปรด้วยวิธี Lasso, Adaptive Lasso และ SCAD สามารถหาค่า $|\hat{S}|$ ได้มากที่สุดและเข้าใกล้ $|S|$ มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วยวิธี Lasso, Adaptive Lasso และ SCAD จึงมีความเหมาะสมที่สุด ที่ $\rho = 0.5$ การคัดกรองตัวแปรด้วยวิธี SCAD สามารถหาค่า $|\hat{S}|$ ได้มากที่สุดและเข้าใกล้ $|S|$ มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วยวิธี SCAD จึงมีความเหมาะสมที่สุด และที่ $\rho = 0.9$ การคัดกรองตัวแปรด้วยวิธี Adaptive Lasso สามารถหาค่า $|\hat{S}|$ ได้มากที่สุดและเข้าใกล้ $|S|$ มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วยวิธี Adaptive Lasso จึงมีความเหมาะสมที่สุด
- เมื่อ $|S|$ มี 40 ค่า เมื่อตัวแปรอิสระมีค่าความสัมพันธ์ที่ $\rho = 0$, 0.5 และ 0.9 การคัดกรองตัวแปรด้วยวิธี Adaptive Lasso สามารถหาค่า $|\hat{S}|$ ได้มากที่สุดและเข้าใกล้ $|S|$ มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วยวิธี Adaptive Lasso จึงมีความเหมาะสมที่สุด
- เมื่อ $|S|$ มี 100 ค่า เมื่อตัวแปรอิสระมีค่าความสัมพันธ์ที่ $\rho = 0$ การคัดกรองตัวแปรด้วยวิธี Adaptive Lasso สามารถหาค่า $|\hat{S}|$ ได้เท่ากันซึ่งมีค่ามากที่สุดและเข้าใกล้ $|S|$ มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วยวิธี Adaptive Lasso จึงมีความเหมาะสมที่สุด และที่ $\rho = 0.5$ และ 0.9 การคัดกรองตัวแปรด้วยวิธี SCAD สามารถหาค่า $|\hat{S}|$ ได้เท่ากันซึ่งมีค่ามากที่สุดและเข้าใกล้ $|S|$ มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วยวิธี SCAD จึงมีความเหมาะสมที่สุด

2.) ที่จำนวนตัวแปรอิสระ (p) เท่ากับ 1000

- เมื่อ $|S|$ มี 20 ค่า เมื่อตัวแปรอิสระมีค่าความสัมพันธ์ที่ $\rho = 0$, 0.5 และ 0.9 การคัดกรองตัวแปรด้วยวิธี SCAD สามารถหาค่า $|\hat{S}|$ ได้มากที่สุดและเข้าใกล้ $|S|$ มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วยวิธี SCAD จึงมีความเหมาะสมที่สุด
- เมื่อ $|S|$ มี 40 ค่า เมื่อตัวแปรอิสระมีค่าความสัมพันธ์ที่ $\rho = 0$, 0.5 และ 0.9 การคัดกรองตัวแปรด้วยวิธี Adaptive Lasso สามารถหาค่า $|\hat{S}|$ ได้มากที่สุดและเข้าใกล้ $|S|$ มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วยวิธี Adaptive Lasso จึงมีความเหมาะสมที่สุด
- เมื่อ $|S|$ มี 100 ค่า เมื่อตัวแปรอิสระมีค่าความสัมพันธ์ที่ $\rho = 0$ และ 0.5 การคัดกรองตัวแปรด้วยวิธี Adaptive Lasso สามารถหาค่า $|\hat{S}|$ ได้มากที่สุดและเข้าใกล้ $|S|$ มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วยวิธี Adaptive Lasso จึงมีความเหมาะสมที่สุด ที่ $\rho = 0.9$ การคัดกรองตัวแปรด้วยวิธี SCAD สามารถหาค่า $|\hat{S}|$ ได้มากที่สุดและเข้าใกล้ $|S|$ มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วยวิธี SCAD จึงมีความเหมาะสมที่สุด

3.) ที่จำนวนตัวแปรอิสระ (p) เท่ากับ 2000

- เมื่อ $|S|$ มี 20 ค่า เมื่อตัวแปรอิสระมีค่าความสัมพันธ์ที่ $\rho = 0, 0.5$ และ 0.9 การคัดกรองตัวแปรด้วยวิธี SCAD สามารถหาค่า $|\hat{S}|$ ได้มากที่สุดและเข้าใกล้ $|S|$ มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วยวิธี SCAD จึงมีความเหมาะสมที่สุด
- เมื่อ $|S|$ มี 40 ค่า เมื่อตัวแปรอิสระมีค่าความสัมพันธ์ที่ $\rho = 0, 0.5$ และ 0.9 การคัดกรองตัวแปรด้วยวิธี Adaptive Lasso สามารถหาค่า $|\hat{S}|$ ได้มากที่สุดและเข้าใกล้ $|S|$ มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วยวิธี Adaptive Lasso จึงมีความเหมาะสมที่สุด
- เมื่อ $|S|$ มี 100 ค่า เมื่อตัวแปรอิสระมีค่าความสัมพันธ์ที่ $\rho = 0$ การคัดกรองตัวแปรด้วยวิธี Adaptive Lasso และ SCAD สามารถหาค่า $|\hat{S}|$ ได้มากที่สุดและเข้าใกล้ $|S|$ มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วยวิธี Adaptive Lasso และ SCAD จึงมีความเหมาะสมที่สุด ที่ $\rho = 0.5$ การคัดกรองตัวแปรด้วย 4 วิธีข้างต้นหาค่า $|\hat{S}|$ ได้เท่ากับ 0 ทั้งหมด นั่นคือไม่มีวิธีการคัดกรองตัวแปรวิธีใดที่หาค่าสัมประสิทธิ์ไม่เท่ากับ 0 จากการทดสอบสมมติฐานได้เลย และที่ $\rho = 0.9$ การคัดกรองตัวแปรด้วย SCAD สามารถหาค่า $|\hat{S}|$ ได้มากที่สุดและเข้าใกล้ $|S|$ มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วยวิธี SCAD จึงมีความเหมาะสมที่สุด

และจากผลในตารางที่ 4.3.2 ยังสามารถสรุปได้อีกว่า

- ในกรณีที่ p มีจำนวนเพิ่มขึ้นจะทำให้ประสิทธิภาพในการหาค่า $|\hat{S}|$ ให้เข้าใกล้ $|S|$ ลดลง
- ในกรณีที่ ρ มีค่าเพิ่มขึ้นจะทำให้ประสิทธิภาพในการหาค่า $|\hat{S}|$ ให้เข้าใกล้ $|S|$ ลดลง
- ในกรณีที่ $|S|$ มีจำนวนเพิ่มขึ้นจะทำให้ประสิทธิภาพในการหาค่า $|\hat{S}|$ ให้เข้าใกล้ $|S|$ ลดลง

และจากตารางที่ 4.3.1 และตารางที่ 4.3.2 จะได้ว่าขนาดของสัมประสิทธิ์ที่ไม่เท่ากับ 0 ที่ขนาดใหญ่ (Large effect Size) จะหาค่า $|\hat{S}|$ เข้าใกล้ $|S|$ ได้ดีกว่าขนาดเล็ก (Small effect Size)

ตารางที่ 4.3.3 แสดงค่าเฉลี่ย(ค่าเบี่ยงเบนมาตรฐาน) ของ FP เมื่อควบคุม FDR ที่ระดับ 0.1 โดยคำนวณจากข้อมูล 100 ชุด กรณีที่ขนาดตัวอย่าง (n) เท่ากับ 200 และขนาดของสัมประสิทธิ์ที่ไม่เท่ากับ 0 มีขนาดเล็ก (Small effect Size)

n : p	|S|	ขนาด(Effect size) ของสัมประสิทธิ์ที่ไม่เท่ากับ 0 มีขนาดเล็ก $(0 < \|\beta\| < 1)$											
		ความสัมพันธ์ (Correlation) ของตัวแปรอิสระ											
		$\rho = 0$				$\rho = 0.5$				$\rho = 0.9$			
		Lasso	Adaptive Lasso	EN	SCAD	Lasso	Adaptive Lasso	EN	SCAD	Lasso	Adaptive Lasso	EN	SCAD
200:400	20	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0.030 (0.171)	0 (0.000)	0.010 (0.100)	0.020 (0.141)	2.490 (1.642)	0 (0.000)	4.580 (2.216)
	40	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0.020 (0.141)	0 (0.000)	0.010 (0.100)	0 (0.000)	0.1440 (1.148)	0 (0.000)	3.020 (1.820)
	100	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0.288 (0.626)	0 (0.000)	1.422 (1.738)
200:1000	20	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0.020 (0.141)	0 (0.000)	0 (0.000)	0.020 (0.141)	3.160 (1.894)	0 (0.000)	3.430 (1.991)
	40	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0.850 (1.157)	0 (0.000)	0.910 (1.334)
	100	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0.230 (0.489)	0 (0.000)	0.090 (0.287)
200:2000	20	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0.010 (0.100)	0 (0.000)	0.010 (0.100)	0.040 (0.196)	2.660 (1.985)	0 (0.000)	1.930 (1.849)
	40	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0.040 (0.196)	0.560 (0.832)	0 (0.000)	0.220 (0.542)
	100	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0.020 (0.141)	0.040 (0.196)	0 (0.000)	0 (0.000)

หมายเหตุ ช่องที่ระบายสี หมายถึง วิธีที่เหมาะสมที่สุดในแต่ละกรณี

จากตารางที่ 4.3.3 ซึ่งแสดงผลของ False Positive โดยเฉลี่ยของข้อมูลจำลองขนาด 100 ระหว่างการคัดกรองตัวแปรด้วยวิธี Lasso, Adaptive Lasso, En และ SCAD โดยเปรียบเทียบกับจำนวนสัมประสิทธิ์ของตัวแบบที่แท้จริง ที่ขนาดของสัมประสิทธิ์ไม่เท่ากับ 0 มีขนาดเล็ก $(0 < |\beta| < 1)$ พบว่า

1.) ที่จำนวนตัวแปรอิสระ (p) เท่ากับ 400, 1000 และ 2000 เมื่อ $\rho = 0$ การคัดกรองตัวแปรด้วยทั้ง 4 วิธี ข้างต้นให้ค่า FP เท่ากับ 0 ทั้งหมด

2.) ที่จำนวนตัวแปรอิสระ (p) เท่ากับ 400

- เมื่อ $|S|$ มี 20 ค่า และตัวแปรอิสระมีค่าความสัมพันธ์ที่ $\rho = 0.5$ การคัดกรองตัวแปรด้วยวิธี Lasso และ EN จะมีค่า FP ต่ำที่สุดและเข้าใกล้ 0 มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วยวิธี Lasso และ EN จึงเหมาะสมที่สุด และ $\rho = 0.5$ การคัดกรองตัวแปรด้วยวิธี EN จะมีค่า FP ต่ำที่สุดและเข้าใกล้ 0 มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วยวิธี EN จึงเหมาะสมที่สุด
- เมื่อ $|S|$ มี 40 ค่า และตัวแปรอิสระมีค่าความสัมพันธ์ที่ $\rho = 0.5$ และ 0.9 การคัดกรองตัวแปรด้วยวิธี Lasso และ EN จะมีค่า FP ต่ำที่สุดและเข้าใกล้ 0 มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วยวิธี Lasso และ EN จึงเหมาะสมที่สุด
- เมื่อ $|S|$ มี 100 ค่า และตัวแปรอิสระมีค่าความสัมพันธ์ที่ $\rho = 0.5$ การคัดกรองตัวแปรด้วยทั้ง 4 วิธีข้างต้นให้ค่า FP เท่ากับ 0 ทั้งหมดและที่ $\rho = 0.9$ การคัดกรองตัวแปรด้วยวิธี Lasso และ EN จะมีค่า FP ต่ำที่สุดและเข้าใกล้ 0 มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วยวิธี Lasso และ EN จึงเหมาะสมที่สุด

3.) ที่จำนวนตัวแปรอิสระ (p) เท่ากับ 1000

- เมื่อ $|S|$ มี 20 ค่า และตัวแปรอิสระมีค่าความสัมพันธ์ที่ $\rho = 0.5$ การคัดกรองตัวแปรด้วยวิธี Lasso, Adaptive Lasso และ EN จะมีค่า FP ต่ำที่สุดและเข้าใกล้ 0 มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วยวิธี Lasso, Adaptive Lasso และ EN จึงเหมาะสมที่สุด และที่ $\rho = 0.9$ การคัดกรองตัวแปรด้วยวิธี EN จะมีค่า FP ต่ำที่สุดและเข้าใกล้ 0 มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วยวิธี EN จึงเหมาะสมที่สุด
- เมื่อ $|S|$ มี 40 และ 100 ค่า และตัวแปรอิสระมีค่าความสัมพันธ์ที่ $\rho = 0.5$ การคัดกรองตัวแปรด้วยทั้ง 4 วิธีข้างต้นให้ค่า FP เท่ากับ 0 ทั้งหมด และที่ $\rho = 0.9$ การคัดกรองตัวแปรด้วยวิธี Lasso และ EN จะมีค่า FP ต่ำที่สุดและเข้าใกล้ 0 มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วยวิธี Lasso และ EN จึงเหมาะสมที่สุด

4.) ที่จำนวนตัวแปรอิสระ (p) เท่ากับ 2000

- เมื่อ $|S|$ มี 20 ค่า และตัวแปรอิสระมีค่าความสัมพันธ์ที่ $\rho = 0.5$ การคัดกรองตัวแปรด้วยวิธี Lasso และ EN จะมีค่า FP ต่ำที่สุดและเข้าใกล้ 0 มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วยวิธี Lasso และ EN จึงเหมาะสมที่สุด และที่ $\rho = 0.9$ การคัดกรองตัวแปรด้วยวิธี

EN จะมีค่า FP ต่ำที่สุดและเข้าใกล้ 0 มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วยวิธี EN จึงเหมาะสมที่สุด

- เมื่อ $|S|$ มี 40 ค่า และตัวแปรอิสระมีค่าความสัมพันธ์ที่ 0.5 การคัดกรองตัวแปรด้วยทั้ง 4 วิธีข้างต้นให้ค่า FP เท่ากับ 0 ทั้งหมด ที่ $\rho = 0.9$ การคัดกรองตัวแปรด้วยวิธี EN จะมีค่า FP ต่ำที่สุดและเข้าใกล้ 0 มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วยวิธี EN จึงเหมาะสมที่สุด
- เมื่อ $|S|$ มี 100 ค่า และตัวแปรอิสระมีค่าความสัมพันธ์ที่ 0.5 การคัดกรองตัวแปรด้วยทั้ง 4 วิธีข้างต้นให้ค่า FP เท่ากับ 0 ทั้งหมด ที่ $\rho = 0.9$ การคัดกรองตัวแปรด้วยวิธี EN และ SCAD จะมีค่า FP ต่ำที่สุดและเข้าใกล้ 0 มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วยวิธี EN และ SCAD จึงเหมาะสมที่สุด

และจากผลในตารางที่ 4.3.3 ยังสามารถสรุปได้อีกว่า

- ในกรณีที่ p มีจำนวนเพิ่มขึ้นจะทำให้ประสิทธิภาพในการหาค่า FP ให้เข้าใกล้ 0 เพิ่มขึ้น
- ในกรณีที่ ρ มีค่าเพิ่มขึ้นจะทำให้ประสิทธิภาพในการหาค่า FP ให้เข้าใกล้ 0 ลดลง
- ในกรณีที่ $|S|$ มีจำนวนเพิ่มขึ้นจะทำให้ประสิทธิภาพในการหาค่า FP ให้เข้าใกล้ 0 เพิ่มขึ้น

ตารางที่ 4.3.4 แสดงค่าเฉลี่ย(ค่าเบี่ยงเบนมาตรฐาน) ของ FP เมื่อควบคุม FDR ที่ระดับ 0.1 โดยคำนวณจากข้อมูล 100 ชุด กรณีที่ขนาดตัวอย่าง (n) เท่ากับ 200 และขนาดของสัมประสิทธิ์ที่ไม่เท่ากับ 0 มีขนาดใหญ่ (Large effect Size)

n : p	|S|	ขนาด(Effect size) ของสัมประสิทธิ์ที่ไม่เท่ากับ 0 มีขนาดใหญ่ $(1 < \|\beta\| < 10)$											
		ความสัมพันธ์ (Correlation) ของตัวแปรอิสระ											
		$\rho = 0$				$\rho = 0.5$				$\rho = 0.9$			
		Lasso	Adaptive Lasso	EN	SCAD	Lasso	Adaptive Lasso	EN	SCAD	Lasso	Adaptive Lasso	EN	SCAD
200:400	20	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0.020 (0.141)	0 (0.000)	0 (0.000)	0.010 (0.100)	2.430 (1.753)	0 (0.000)	9.280 (3.197)
	40	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0.010 (0.100)	0 (0.000)	0.010 (0.100)	0.500 (2.222)	3.400 (5.389)	0 (0.000)	6.620 (4.044)
	100	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0.200 (0.568)	0 (0.000)	2.100 (1.898)
200:1000	20	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0.030 (0.171)	0 (0.000)	0 (0.000)	0 (0.000)	1.540 (0.138)	0 (0.000)	1.240 (0.080)
	40	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)
	100	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0.100 (0.301)	0 (0.000)	0.180 (0.479)
200:2000	20	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)
	40	0 (0.000)	0.010 (0.100)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0.840 (1.721)	0 (0.000)	0.530 (1.209)
	100	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)	0.040 (0.196)

หมายเหตุ ช่องที่ระบายสี หมายถึง วิธีที่เหมาะสมที่สุดในแต่ละกรณี

จากตารางที่ 4.3.4 ซึ่งแสดงผลของ False Positive โดยเฉลี่ยของข้อมูลจำลองขนาด 100 ระหว่างการคัดกรองตัวแปรด้วยวิธี Lasso, Adaptive Lasso, En และ SCAD โดยเปรียบเทียบกับจำนวนสัมประสิทธิ์ของตัวแบบที่แท้จริง ที่ขนาด(Effect Size) ของสัมประสิทธิ์ไม่เท่ากับ 0 เป็น Large Size $(1 < |\beta| < 10)$ พบว่า

1.) ที่จำนวนตัวแปรอิสระ (p) เท่ากับ 400

- เมื่อ $|S|$ มี 20 ค่า และตัวแปรอิสระมีค่าความสัมพันธ์ที่ $\rho = 0$ การคัดกรองตัวแปรด้วยทั้ง 4 วิธีข้างต้นให้ค่า FP เท่ากับ 0 ทั้งหมด ที่ $\rho = 0.5$ การคัดกรองตัวแปรด้วยวิธี Lasso, EN และ SCAD จะมีค่า FP ต่ำที่สุดและเข้าใกล้ 0 มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วยวิธี Lasso, EN และ SCAD จึงเหมาะสมที่สุด และ $\rho = 0.9$ การคัดกรองตัวแปรด้วยวิธี EN จะมีค่า FP ต่ำที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วยวิธี EN จึงเหมาะสมที่สุด
- เมื่อ $|S|$ มี 40 ค่า และตัวแปรอิสระมีค่าความสัมพันธ์ที่ $\rho = 0$ การคัดกรองตัวแปรด้วยทั้ง 4 วิธีข้างต้นให้ค่า FP เท่ากับ 0 ทั้งหมด ที่ $\rho = 0.5$ การคัดกรองตัวแปรด้วยวิธี Lasso และ EN จะมีค่า FP ต่ำที่สุดและเข้าใกล้ 0 มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วยวิธี Lasso และ EN จึงเหมาะสมที่สุดและ $\rho = 0.9$ การคัดกรองตัวแปรด้วยวิธี EN จะมีค่า FP ต่ำที่สุดและเข้าใกล้ 0 มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วยวิธี EN จึงเหมาะสมที่สุด
- เมื่อ $|S|$ มี 100 ค่า และตัวแปรอิสระมีค่าความสัมพันธ์ที่ $\rho = 0$ และ 0.5 การคัดกรองตัวแปรด้วยทั้ง 4 วิธีข้างต้นให้ค่า FP เท่ากับ 0 ทั้งหมด และที่ $\rho = 0.9$ การคัดกรองตัวแปรด้วยวิธี Lasso และ EN จะมีค่า FP ต่ำที่สุดและเข้าใกล้ 0 มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วยวิธี Lasso และ EN จึงเหมาะสมที่สุด

2.) ที่จำนวนตัวแปรอิสระ (p) เท่ากับ 1000

- เมื่อ $|S|$ มี 20 ค่า และตัวแปรอิสระมีค่าความสัมพันธ์ที่ $\rho = 0$ การคัดกรองตัวแปรด้วยทั้ง 4 วิธีข้างต้นให้ค่า FP เท่ากับ 0 ทั้งหมด ที่ $\rho = 0.5$ การคัดกรองตัวแปรด้วยวิธี Lasso, EN และ SCAD จะมีค่า FP ต่ำที่สุดและเข้าใกล้ 0 มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วยวิธี Lasso, EN และ SCAD จึงเหมาะสมที่สุด และ $\rho = 0.9$ การคัดกรองตัวแปรด้วยวิธี Lasso และ EN จะมีค่า FP ต่ำที่สุดและเข้าใกล้ 0 มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วยวิธี Lasso และ EN จึงเหมาะสมที่สุด
- เมื่อ $|S|$ มี 40 ค่า และตัวแปรอิสระมีค่าความสัมพันธ์ที่ $\rho = 0$, 0.5 และ 0.9 การคัดกรองตัวแปรด้วยทั้ง 4 วิธีข้างต้นให้ค่า FP เท่ากับ 0 ทั้งหมด
- เมื่อ $|S|$ มี 100 ค่า และตัวแปรอิสระมีค่าความสัมพันธ์ที่ $\rho = 0$ และ 0.5 การคัดกรองตัวแปรด้วยทั้ง 4 วิธีข้างต้นให้ค่า FP เท่ากับ 0 ทั้งหมด และที่ $\rho = 0.9$ การคัดกรองตัวแปรด้วยวิธี Lasso และ EN จะมีค่า FP ต่ำที่สุดและเข้าใกล้ 0 มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วยวิธี Lasso และ EN จึงเหมาะสมที่สุด

3.) ที่จำนวนตัวแปรอิสระ (p) เท่ากับ 2000

- เมื่อ $|S|$ มี 20 ค่า และตัวแปรอิสระมีค่าความสัมพันธ์ที่ $\rho = 0$, 0.5 และ 0.9 การคัดกรองตัวแปรด้วยทั้ง 4 วิธีข้างต้นให้ค่า FP เท่ากับ 0 ทั้งหมด
- เมื่อ $|S|$ มี 40 ค่า และตัวแปรอิสระมีค่าความสัมพันธ์ที่ 0 การคัดกรองตัวแปรด้วยวิธี Lasso, EN และ SCAD จะมีค่า FP ต่ำที่สุดและเข้าใกล้ 0 มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วยวิธี Lasso, EN และ SCAD จึงเหมาะสมที่สุด ที่ $\rho = 0.5$ การคัดกรองตัวแปรด้วยทั้ง 4 วิธีข้างต้นให้ค่า FP เท่ากับ 0 ทั้งหมด และที่ $\rho = 0.9$ การคัดกรองตัวแปรด้วยวิธี Lasso และ EN จะมีค่า FP ต่ำที่สุดและเข้าใกล้ 0 มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วยวิธี Lasso และ EN จึงเหมาะสมที่สุด
- เมื่อ $|S|$ มี 100 ค่า และตัวแปรอิสระมีค่าความสัมพันธ์ที่ 0 และ 0.5 การคัดกรองตัวแปรด้วยทั้ง 4 วิธีข้างต้นให้ค่า FP เท่ากับ 0 ทั้งหมด ที่ $\rho = 0.9$ การคัดกรองตัวแปรด้วยวิธี Lasso, Adaptive Lasso และ SCAD จะมีค่า FP ต่ำที่สุดและเข้าใกล้ 0 มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วยวิธี Lasso, Adaptive Lasso และ SCAD จึงเหมาะสมที่สุด

และจากผลในตารางที่ 4.3.3 ยังสามารถสรุปได้อีกว่า

- ในกรณีที่ p มีจำนวนเพิ่มขึ้นจะทำให้ประสิทธิภาพในการหาค่า FP ให้เข้าใกล้ 0 เพิ่มขึ้น
- ในกรณีที่ ρ มีค่าเพิ่มขึ้นจะทำให้ประสิทธิภาพในการหาค่า FP ให้เข้าใกล้ 0 ลดลง
- ในกรณีที่ $|S|$ มีจำนวนเพิ่มขึ้นจะทำให้ประสิทธิภาพในการหาค่า FP ให้เข้าใกล้ 0 เพิ่มขึ้น

และจากตารางที่ 4.3.3 และตารางที่ 4.3.4 จะได้ว่าขนาดของสัมประสิทธิ์ที่ไม่เท่ากับ 0 ที่ขนาดใหญ่ (Large effect Size) จะมีประสิทธิภาพในการหาค่า FP เข้าใกล้ 0 ได้ดีกว่าขนาดเล็ก (Small effect Size)

ตารางที่ 4.3.5 แสดงค่าเฉลี่ย(ค่าเบี่ยงเบนมาตรฐาน) ของ FN เมื่อควบคุม FDR ที่ระดับ 0.1 โดยคำนวณจากข้อมูล 100 ชุด กรณีที่ขนาดตัวอย่าง (n) เท่ากับ 200 และขนาดของสัมประสิทธิ์ที่ไม่เท่ากับ 0 มีขนาดเล็ก (Small effect Size)

n : p	|S|	ขนาด(Effect size) ของสัมประสิทธิ์ที่ไม่เท่ากับ 0 มีขนาดเล็ก $(0 < \|\beta\| < 1)$											
		ความสัมพันธ์ (Correlation) ของตัวแปรอิสระ											
		$\rho = 0$				$\rho = 0.5$				$\rho = 0.9$			
		Lasso	Adaptive Lasso	EN	SCAD	Lasso	Adaptive Lasso	EN	SCAD	Lasso	Adaptive Lasso	EN	SCAD
200:400	20	10.880	8.920	17.860	8.630	11.830	9.720	18.560	9.440	17.750	11.660	19.980	11.690
		(1.793)	(2.355)	(2.160)	(2.359)	(2.122)	(2.357)	(1.827)	(2.400)	(1.737)	(2.207)	(0.141)	(1.978)
	40	37.880	33.560	40.000	35.960	38.720	33.160	39.990	36.120	39.860	32.270	40.000	33.170
		(2.811)	(4.304)	(0.000)	(3.314)	(1.969)	(4.263)	(0.100)	(3.002)	(0.512)	(3.384)	(0.000)	(3.241)
	100	99.980	99.960	100.000	99.950	100.000	99.930	100.000	99.910	100.000	98.111	100.000	96.422
		(0.141)	(0.196)	(0.000)	(0.219)	(0.000)	(0.256)	(0.000)	(0.287)	(0.000)	(2.025)	(0.000)	(2.957)
200:1000	20	12.170	10.550	19.549	10.200	13.920	12.010	19.840	11.840	17.640	11.870	20.000	12.740
		(2.482)	(2.614)	(0.903)	(2.522)	(2.634)	(2.999)	(0.486)	(2.740)	(1.755)	(2.372)	(0.000)	(0.000)
	40	39.440	38.840	40.000	38.980	39.600	38.520	39.990	38.710	39.890	36.860	40.000	37.940
		(0.782)	(1.454)	(0.000)	(1.197)	(1.024)	(2.148)	(0.100)	(2.046)	(0.423)	(2.895)	(0.000)	(2.102)
	100	99.990	99.990	100.000	99.990	99.990	100.000	99.980	99.980	100.000	99.390	100.000	99.500
		(0.100)	(0.100)	(0.000)	(0.100)	(0.100)	(0.000)	(0.141)	(0.141)	(0.000)	(0.983)	(0.000)	(0.784)
200:2000	20	14.410	13.000	19.710	12.710	15.550	14.150	19.810	14.150	18.030	13.550	20.000	14.400
		(2.437)	(2.704)	(0.640)	(2.861)	(2.324)	(2.768)	(0.580)	(2.865)	(1.904)	(2.571)	(0.000)	(2.662)
	40	39.730	39.660	40.000	39.600	39.780	39.530	39.990	39.580	39.960	38.780	40.000	39.480
		(0.649)	(0.639)	(0.000)	(0.724)	(0.542)	(1.077)	(0.100)	(0.986)	(0.196)	(1.411)	(0.000)	(0.947)
	100	100.000	100.000	100.000	100.000	99.970	99.970	100.000	99.970	100.000	99.960	100.000	99.960
		(0.000)	(0.000)	(0.000)	(0.000)	(0.171)	(0.171)	(0.000)	(0.171)	(0.000)	(0.196)	(0.000)	(0.196)

หมายเหตุ ช่องที่ระบายสี หมายถึง วิธีที่เหมาะสมที่สุดในแต่ละกรณี

จากตารางที่ 4.3.5 ซึ่งแสดงผลของ False Negative โดยเฉลี่ยของข้อมูลจำลองขนาด 200 ระหว่างการคัดกรองตัวแปรด้วยวิธี Lasso, Adaptive Lasso, En และ SCAD โดยเปรียบเทียบกับจำนวนสัมประสิทธิ์ของตัวแบบที่แท้จริง ที่ขนาดของสัมประสิทธิ์ไม่เท่ากับ 0 มีขนาดเล็ก $(0 < |\beta| < 1)$ พบว่า

1.) ที่จำนวนตัวแปรอิสระ (p) เท่ากับ 400

- เมื่อ $|S|$ มี 20 ค่า และตัวแปรอิสระมีค่าความสัมพันธ์ที่ $\rho = 0$ และ 0.5 การคัดกรองตัวแปรด้วยวิธี SCAD จะมีค่า FN ต่ำที่สุดและเข้าใกล้ 0 มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วย SCAD จึงเหมาะสมที่สุด และที่ $\rho = 0.9$ การคัดกรองตัวแปรด้วยวิธี Adaptive Lasso จะมีค่า FN ต่ำที่สุดและเข้าใกล้ 0 มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วยวิธี Adaptive Lasso จึงเหมาะสมที่สุด
- เมื่อ $|S|$ มี 40 ค่า และตัวแปรอิสระมีค่าความสัมพันธ์ที่ $\rho = 0$, 0.5 และ 0.9 การคัดกรองตัวแปรด้วยวิธี Adaptive Lasso จะมีค่า FN ต่ำที่สุดและเข้าใกล้ 0 มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วยวิธี Adaptive Lasso จึงเหมาะสมที่สุด
- เมื่อ $|S|$ มี 100 ค่า และตัวแปรอิสระมีค่าความสัมพันธ์ที่ $\rho = 0$, 0.5 และ 0.9 การคัดกรองตัวแปรด้วยวิธี SCAD จะมีค่า FN ต่ำที่สุดและเข้าใกล้ 0 มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วยวิธี SCAD จึงเหมาะสมที่สุด

2.) ที่จำนวนตัวแปรอิสระ (p) เท่ากับ 1000

- เมื่อ $|S|$ มี 20 ค่า และตัวแปรอิสระมีค่าความสัมพันธ์ที่ $\rho = 0$ และ 0.5 การคัดกรองตัวแปรด้วยวิธี SCAD จะมีค่า FN ต่ำที่สุดและเข้าใกล้ 0 มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วย SCAD จึงเหมาะสมที่สุด และที่ $\rho = 0.9$ การคัดกรองตัวแปรด้วยวิธี Adaptive Lasso จะมีค่า FN ต่ำที่สุดและเข้าใกล้ 0 มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วยวิธี Adaptive Lasso จึงเหมาะสมที่สุด
- เมื่อ $|S|$ มี 40 ค่า และตัวแปรอิสระมีค่าความสัมพันธ์ที่ $\rho = 0$, 0.5 และ 0.9 การคัดกรองตัวแปรด้วยวิธี Adaptive Lasso จะมีค่า FN ต่ำที่สุดและเข้าใกล้ 0 มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วยวิธี Adaptive Lasso จึงเหมาะสมที่สุด
- เมื่อ $|S|$ มี 100 ค่า และตัวแปรอิสระมีค่าความสัมพันธ์ที่ $\rho = 0$ การคัดกรองตัวแปรด้วยวิธี Lasso, Adaptive Lasso และ SCAD จะมีค่า FN ต่ำที่สุดและเข้าใกล้ 0 มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วยวิธี Lasso, Adaptive Lasso และ SCAD จึงเหมาะสมที่สุด ที่ $\rho = 0.5$ การคัดกรองตัวแปรด้วยวิธี EN และ SCAD จะมีค่า FN ต่ำที่สุดและเข้าใกล้ 0 มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วยวิธี EN และ SCAD จึงเหมาะสมที่สุด และที่ $\rho = 0.9$ การคัดกรองตัวแปรด้วยวิธี Adaptive Lasso จะมีค่า FN ต่ำที่สุดและเข้าใกล้ 0 มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วยวิธี Adaptive Lasso จึงเหมาะสมที่สุด

3.) ที่จำนวนตัวแปรอิสระ (p) เท่ากับ 2000

- เมื่อ $|S|$ มี 20 ค่า และตัวแปรอิสระมีค่าความสัมพันธ์ที่ $\rho = 0$ การคัดกรองตัวแปรด้วยวิธี SCAD จะมีค่า FN ต่ำที่สุดและเข้าใกล้ 0 มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วยวิธี SCAD จึงเหมาะสมที่สุด และที่ $\rho = 0.5$ และ 0.9 การคัดกรองตัวแปรด้วยวิธี Adaptive Lasso จะมีค่า FN ต่ำที่สุดและเข้าใกล้ 0 มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วย Adaptive Lasso จึงเหมาะสมที่สุด
- เมื่อ $|S|$ มี 40 ค่า และตัวแปรอิสระมีค่าความสัมพันธ์ที่$\rho = 0$ และ 0.9 การคัดกรองตัวแปรด้วยวิธี SCAD จะมีค่า FN ต่ำที่สุดและเข้าใกล้ 0 มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วยวิธี SCAD จึงเหมาะสมที่สุด และที่ $\rho = 0.5$ การคัดกรองตัวแปรด้วยวิธี Adaptive Lasso จะมีค่า FN ต่ำที่สุดและเข้าใกล้ 0 มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วย Adaptive Lasso จึงเหมาะสมที่สุด
- เมื่อ $|S|$ มี 100 ค่า และตัวแปรอิสระมีค่าความสัมพันธ์ที่ $\rho = 0$ การคัดกรองตัวแปรทั้ง 4 วิธีข้างต้นให้ค่า FN เท่ากันทั้งหมด ที่ $\rho = 0.5$ การคัดกรองตัวแปรด้วยวิธี Lasso, Adaptive Lasso และ SCAD จะมีค่า FN ต่ำที่สุดและเข้าใกล้ 0 มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วย Lasso, Adaptive Lasso และ SCAD จึงเหมาะสมที่สุด และที่ $\rho = 0.9$ การคัดกรองตัวแปรด้วยวิธี Adaptive Lasso และ SCAD จะมีค่า FN ต่ำที่สุดและเข้าใกล้ 0 มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรวิธี Adaptive Lasso และ SCAD จึงเหมาะสมที่สุด

และจากผลในตารางที่ 4.3.5 ยังสามารถสรุปได้อีกว่า

- ในกรณีที่ p มีจำนวนเพิ่มขึ้นจะทำให้ประสิทธิภาพในการหาค่า FN ให้เข้าใกล้ 0 ลดลง
- ในกรณีที่ ρ มีค่าเพิ่มขึ้นจะทำให้ประสิทธิภาพในการหาค่า FN ให้เข้าใกล้ 0 ลดลง
- ในกรณีที่ $|S|$ มีจำนวนเพิ่มขึ้นจะทำให้ประสิทธิภาพในการหาค่า FN ให้เข้าใกล้ 0 ลดลง

ตารางที่ 4.3.6 แสดงค่าเฉลี่ย(ค่าเบี่ยงเบนมาตรฐาน) ของ FN เมื่อควบคุม FDR ที่ระดับ 0.1 โดยคำนวณจากข้อมูล 100 ชุด กรณีที่ขนาดตัวอย่าง (n) เท่ากับ 200 และขนาดของสัมประสิทธิ์ที่ไม่เท่ากับ 0 มีขนาดใหญ่ (Large effect Size)

n : p	|S|	ขนาด(Effect size) ของสัมประสิทธิ์ที่ไม่เท่ากับ 0 มีขนาดใหญ่ $(1 < \|\beta\| < 10)$											
		ความสัมพันธ์ (Correlation) ของตัวแปรอิสระ											
		$\rho = 0$				$\rho = 0.5$				$\rho = 0.9$			
		Lasso	Adaptive Lasso	EN	SCAD	Lasso	Adaptive Lasso	EN	SCAD	Lasso	Adaptive Lasso	EN	SCAD
200:400	20	0	0	0.940	0	0.050	0	1.430	0	1.400	0.450	3.040	1.570
		(0.000)	(0.000)	(1.873)	(0.000)	(0.219)	(0.000)	(2.868)	(0.000)	(1.504)	(0.701)	(3.314)	(1.357)
	40	34.260	27.410	39.860	34.970	34.470	25.690	40.000	34.230	37.020	21.260	40.000	28.000
		(6.299)	(7.845)	(0.804)	(5.255)	(6.009)	(7.306)	(0.000)	(4.836)	(4.383)	(7.310)	(0.000)	(5.041)
	100	99.990	99.970	100.000	99.980	100.000	99.960	100.000	99.920	99.980	98.740	100.000	95.260
		(0.100)	(0.171)	(0.000)	(0.141)	(0.000)	(0.242)	(0.000)	(0.307)	(0.141)	(1.851)	(0.000)	(3.183)
200:1000	20	2.300	2.120	14.440	0.020	2.530	1.950	14.140	0	2.270	2.600	13.880	0.560
		(2.713)	(1.838)	(3.924)	(0.141)	(2.341)	(1.754)	(4.722)	(0.000)	(2.436)	(0.138)	(4.137)	(0.890)
	40	39.400	38.080	40.000	38.460	39.460	38.120	40.000	38.610	39.470	38.240	40.000	38.690
		(1.385)	(2.501)	(0.000)	(2.090)	(1.395)	(3.127)	(0.000)	(2.024)	(1.213)	(2.882)	(0.000)	(2.356)
	100	99.990	99.970	100.000	99.980	100.000	99.960	100.000	99.980	100.000	99.720	100.000	99.580
		(0.100)	(0.171)	(0.000)	(0.141)	(0.000)	(0.196)	(0.000)	(0.141)	(0.000)	(0.533)	(0.000)	(0.830)
200:2000	20	8.070	7.170	17.420	0.610	8.330	6.960	18.830	1.110	6.690	6.520	18.210	1.230
		(3.539)	(3.104)	(2.563)	(2.824)	(3.149)	(3.437)	(1.741)	(3.845)	(3.030)	(3.219)	(2.379)	(3.175)
	40	39.740	39.260	39.990	39.400	39.770	39.290	40.000	39.500	39.880	38.450	40.000	38.930
		(0.629)	(1.673)	(0.100)	(1.370)	(0.617)	(1.273)	(0.000)	(1.010)	(0.408)	(2.392)	(0.000)	(1.854)
	100	100.000	99.990	100.000	99.990	100.000	100.000	100.000	100.000	100.000	99.960	100.000	99.980
		(0.000)	(0.100)	(0.000)	(0.100)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.196)	(0.000)	(0.141)

หมายเหตุ ช่องที่ระบายสี หมายถึง วิธีที่เหมาะสมที่สุดในแต่ละกรณี

จากตารางที่ 4.3.6 ซึ่งแสดงผลของ False Negative โดยเฉลี่ยของข้อมูลจำลองขนาด 200 ระหว่างการคัดกรองตัวแปรด้วยวิธี Lasso, Adaptive Lasso, En และ SCAD โดยเปรียบเทียบกับจำนวนสัมประสิทธิ์ของตัวแบบที่แท้จริง ที่ขนาดของสัมประสิทธิ์ไม่เท่ากับ 0 มีขนาดใหญ่ $(1 < |\beta| < 10)$ พบว่า

1.) ที่จำนวนตัวแปรอิสระ (p) เท่ากับ 400

- เมื่อ $|S|$ มี 20 ค่า และตัวแปรอิสระมีค่าความสัมพันธ์ที่ $\rho = 0$ การคัดกรองตัวแปรด้วยวิธี Lasso, Adaptive Lasso และ SCAD จะมีค่า FN ต่ำที่สุดและเข้าใกล้ 0 มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรวิธี Lasso, Adaptive Lasso และ SCAD จึงเหมาะสมที่สุด ที่ $\rho = 0.5$ การคัดกรองตัวแปรด้วยวิธี Adaptive Lasso และ SCAD จะมีค่า FN ต่ำที่สุดและเข้าใกล้ 0 มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วยวิธี Adaptive Lasso และ SCAD จึงเหมาะสมที่สุด และที่ $\rho = 0.9$ การคัดกรองตัวแปรด้วยวิธี Adaptive Lasso จะมีค่า FN ต่ำที่สุดและเข้าใกล้ 0 มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วยวิธี Adaptive Lasso จึงเหมาะสมที่สุด
- เมื่อ $|S|$ มี 40 ค่า และตัวแปรอิสระมีค่าความสัมพันธ์ที่$\rho = 0, 0.5$ และ 0.9 การคัดกรองตัวแปรด้วยวิธี Adaptive Lasso จะมีค่า FN ต่ำที่สุดและเข้าใกล้ 0 มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วยวิธี Adaptive Lasso จึงเหมาะสมที่สุด
- เมื่อ $|S|$ มี 100 ค่า และตัวแปรอิสระมีค่าความสัมพันธ์ที่ $\rho = 0$ การคัดกรองตัวแปรด้วยวิธี Adaptive Lasso จะมีค่า FN ต่ำที่สุดและเข้าใกล้ 0 มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วยวิธี Adaptive Lasso จึงเหมาะสมที่สุด และที่ $\rho = 0.5$ และ 0.9 การคัดกรองตัวแปรด้วยวิธี SCAD จะมีค่า FN ต่ำที่สุดและเข้าใกล้ 0 มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วยวิธี SCAD จึงเหมาะสมที่สุด

2.) ที่จำนวนตัวแปรอิสระ (p) เท่ากับ 1000

- เมื่อ $|S|$ มี 20 ค่า และตัวแปรอิสระมีค่าความสัมพันธ์ที่ $\rho = 0, 0.5$ และ 0.9 การคัดกรองตัวแปรด้วยวิธี SCAD จะมีค่า FN ต่ำที่สุดและเข้าใกล้ 0 มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วย SCAD จึงเหมาะสมที่สุด
- เมื่อ $|S|$ มี 40 ค่า และตัวแปรอิสระมีค่าความสัมพันธ์ที่ $\rho = 0, 0.5$ และ 0.9 การคัดกรองตัวแปรด้วยวิธี Adaptive Lasso จะมีค่า FN ต่ำที่สุดและเข้าใกล้ 0 มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วยวิธี Adaptive Lasso จึงเหมาะสมที่สุด
- เมื่อ $|S|$ มี 100 ค่า และตัวแปรอิสระมีค่าความสัมพันธ์ที่ $\rho = 0$ และ 0.5 การคัดกรองตัวแปรด้วยวิธี Adaptive Lasso จะมีค่า FN ต่ำที่สุดและเข้าใกล้ 0 มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วยวิธี Adaptive Lasso จึงเหมาะสมที่สุด ที่ $\rho = 0.9$ การคัดกรองตัวแปรด้วยวิธี SCAD จะมีค่า FN ต่ำที่สุดและเข้าใกล้ 0 มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วยวิธี SCAD จึงเหมาะสมที่สุด

3.) ที่จำนวนตัวแปรอิสระ (p) เท่ากับ 2000

- เมื่อ $|S|$ มี 20 ค่า และตัวแปรอิสระมีค่าความสัมพันธ์ที่ $\rho = 0, 0.5$ และ 0.9 การคัดกรองตัวแปรด้วยวิธี SCAD จะมีค่า FN ต่ำที่สุดและเข้าใกล้ 0 มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วย SCAD จึงเหมาะสมที่สุด
- เมื่อ $|S|$ มี 40 ค่า และตัวแปรอิสระมีค่าความสัมพันธ์ที่ $\rho = 0, 0.5$ และ 0.9 การคัดกรองตัวแปรด้วยวิธี Adaptive Lasso จะมีค่า FN ต่ำที่สุดและเข้าใกล้ 0 มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรด้วยวิธี Adaptive Lasso จึงเหมาะสมที่สุด
- เมื่อ $|S|$ มี 100 ค่า และตัวแปรอิสระมีค่าความสัมพันธ์ที่ $\rho = 0$ การคัดกรองตัวแปรด้วยวิธี Adaptive Lasso และ SCAD จะมีค่า FN ต่ำที่สุดและเข้าใกล้ 0 มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรวิธี Adaptive Lasso และ SCAD จึงเหมาะสมที่สุด ที่ $\rho = 0.5$ การคัดกรองตัวแปรทั้ง 4 วิธีข้างต้นให้ค่า FN เท่ากันทั้งหมด และที่ $\rho = 0.9$ การคัดกรองตัวแปรด้วยวิธี Adaptive Lasso จะมีค่า FN ต่ำที่สุดและเข้าใกล้ 0 มากที่สุด ดังนั้นในกรณีนี้การคัดกรองตัวแปรวิธี Adaptive Lasso จึงเหมาะสมที่สุด

และจากผลในตารางที่ 4.3.6 ยังสามารถสรุปได้อีกว่า

- ในกรณีที่ p มีจำนวนเพิ่มขึ้นจะทำให้ประสิทธิภาพในการหาค่า FN ให้เข้าใกล้ 0 ลดลง
- ในกรณีที่ ρ มีค่าเพิ่มขึ้นจะทำให้ประสิทธิภาพในการหาค่า FN ให้เข้าใกล้ 0 ลดลง
- ในกรณีที่ $|S|$ มีจำนวนเพิ่มขึ้นจะทำให้ประสิทธิภาพในการหาค่า FN ให้เข้าใกล้ 0 ลดลง

และจากตารางที่ 4.2.5 และตารางที่ 4.2.6 จะได้ว่าขนาดของสัมประสิทธิ์ที่ไม่เท่ากับ 0 ที่ขนาดใหญ่ (Large effect Size) จะหาค่า FN ได้ดีกว่าขนาดเล็ก (Small effect Size)

จากตารางที่ 4.3.1 – 4.3.6 เมื่อพิจารณาจากค่า $|\hat{S}|$, FP และ FN จะได้ว่าเมื่อขนาดตัวอย่าง(n) เท่ากับ 200 ค่าของ $|\hat{S}|$ และ FN จะไปในทิศทางเดียวกัน นั่นคือการคัดกรองตัวแปรด้วยวิธี Adaptive Lasso และวิธี SCAD จะเหมาะสมมากที่สุดและมีอำนาจการทดสอบมากที่สุด แต่จากค่าของ FP จะได้วิธี Lasso และวิธี EN ที่เหมาะสม นั่นแสดงให้เห็นว่าวิธี Lasso และวิธี EN มีประสิทธิภาพในการคัดเลือกตัวแปรและอำนาจในการทดสอบน้อยกว่าวิธี Adaptive Lasso และวิธี SCAD

บทที่ 5

สรุปผลการวิจัยและข้อเสนอแนะ

การศึกษาเปรียบเทียบประสิทธิภาพของวิธีการคัดเลือกตัวแปรและประมาณค่าสัมประสิทธิ์ระหว่างวิธี Lasso วิธี Adaptive Lasso วิธี EN และวิธี SCAD โดยจะพิจารณาแยกตามขนาดตัวอย่างเป็น 10, 100 และ 200 และอัตราส่วนระหว่างขนาดของตัวอย่างและจำนวนตัวแปรอิสระเป็น 1:2, 1:5 และ 1:10 โดยมีเกณฑ์ที่ใช้ในการพิจาณาเปรียบเทียบประสิทธิภาพของแต่ละวิธีจากค่าความผิดพลาดในการตรวจจับเชิงบวก (False Positive : FP) ความผิดพลาดในการตรวจจับเชิงลบ (False Negative : FN) และจำนวนของสัมประสิทธิ์ของตัวแปรอิสระที่มีค่าไม่เท่ากับ 0 จากการทดสอบสมมติฐานโดยเฉลี่ย เมื่อควบคุม FDR ที่ระดับ 0.1 โดยสรุปผลการวิจัยได้ดังนี้

5.1 สรุปผลการวิจัย

5.1.1 แบ่งผลการวิจัยออกเป็น 3 ส่วน โดยพิจารณาตามขนาดของตัวอย่าง ดังนี้

<u>ส่วนที่ 1</u> ผลการเปรียบเทียบความผิดพลาดในการตรวจจับเชิงบวก (FP) ความผิดพลาดในการตรวจจับเชิงลบ (FN) และจำนวนของสัมประสิทธิ์ของตัวแปรอิสระที่มีค่าไม่เท่ากับ 0 จากการทดสอบสมมติฐาน เมื่อควบคุม FDR ที่ระดับ 0.1 ของข้อมูลจำลองขนาด 10 พบว่า

ตารางที่ 5.1.1 แสดงวิธีการคัดกรองตัวแปรที่เหมาะสมที่สุด เมื่อพิจารณาจากค่า $|\hat{S}|$, FP และ FN ระหว่างวิธี Lasso, Adaptive Lasso, EN และ SCAD จากการวิเคราะห์ขนาดตัวอย่าง (n) เท่ากับ 10 โดยจำแนกตามอัตราส่วน ขนาดตัวอย่างต่อจำนวนตัวแปร (n:p) , ขนาด(Effect size) ของสัมประสิทธิ์ที่ไม่เท่ากับ 0 , จำนวนสัมประสิทธิ์ (β) ที่ไม่เท่ากับ 0 และความสัมพันธ์ (Correlation) ของตัวแปรอิสระ

n : p	|S|	ความสัมพันธ์ (Correlation) ของตัวแปรอิสระ					
		$\rho = 0$	$\rho = 0.5$	$\rho = 0.9$	$\rho = 0$	$\rho = 0.5$	$\rho = 0.9$
		Small Size $(0 < \|\beta\| < 1)$			Large Size $(1 < \|\beta\| < 10)$		
พิจารณาจากค่า $\|\hat{S}\|$							
10:20	1	-	-	-	AL	AL	-
	2	-	-	AL+SC	AL	AL	-
	5	-	-	AL	-	-	SC
10:50	1	-	-	-	AL	AL	-
	2	-	-	-	AL	AL	-
	5	-	-	-	-	-	-
10:100	1	-	-	-	AL	AL	-
	2	-	-	-	L+AL	AL+SC	-
	5	-	-	-	-	-	AL+SC
พิจารณาจากค่า FP							
10:20	1	-	-	-	L	-	-
	2	-	-	L	-	-	-
	5	-	-	-	-	-	-
10:50	1	-	-	-	-	-	-
	2	-	-	-	-	-	-
	5	-	-	-	-	-	-
10:100	1	-	-	-	-	-	-
	2	-	-	-	-	-	-
	5	-	-	-	L	-	-
พิจารณาจากค่า FN							
10:20	1	-	-	-	AL	AL	AL
	2	-	-	-	AL	AL	AL
	5	-	-	AL	-	-	SC
10:50	1	-	-	-	AL	AL	AL
	2	-	-	-	AL	AL	AL
	5	-	-	-	-	-	-
10:100	1	-	-	-	AL	AL	AL
	2	-	-	-	L+AL	AL+SC	AL
	5	-	-	-	-	-	-

หมายเหตุ

L หมายถึง วิธี LASSO

AL หมายถึง วิธี Adaptive Lasso

SC หมายถึง วิธี SCAD

จากตารางที่ 5.1 สามารถสรุปผลได้ว่า เมื่อขนาดตัวอย่าง(n) เท่ากับ 10 พิจารณาจากค่า $|\hat{S}|$,FP และ FN โดยค่าของ $|\hat{S}|$ และ FN จะได้วิธีที่เหมาะสมในการคัดกรองตัวแปรคล้ายกัน นั่นคือการคัดกรองตัวแปรด้วยวิธี Adaptive Lasso จะเหมาะสมมากที่สุด โดยทุกวิธีจะสามารถทำงานได้ดีในกรณีที่จำนวนของสัมประสิทธิ์ของตัวแปรอิสระในตัวแบบที่แท้จริงมีค่าไม่เท่ากับ 0 มีจำนวนน้อยๆ แต่จากค่าของ FP การคัดกรองตัวแปรด้วยวิธี Lasso จะเหมาะสมที่สุด โดยที่วิธี Lasso จะเลือกตัวแปรอิสระเข้ามาในการคัดกรองน้อยกว่าความเป็นจริง แต่ตัวแปรที่วิธี Lasso เลือกเข้ามานั้นจะเป็นตัวที่ถูกต้องทั้งหมด นั่นแสดงให้เห็นว่าวิธี Lasso มี ประสิทธิภาพในการคัดเลือกตัวแปรและมีอำนาจในการทดสอบน้อยกว่าวิธี Adaptive Lasso และวิธี SCAD และในกรณีที่ขนาดของสัมประสิทธิ์ที่ไม่เท่ากับ 0 เป็นขนาดเล็ก การคัดกรองจากทั้ง 4 วิธีข้างต้นจะมีประสิทธิภาพเท่าๆกัน

ส่วนที่ 2 ผลเปรียบเทียบความผิดพลาดในการตรวจจับเชิงบวก (FP) ความผิดพลาดในการตรวจจับเชิงลบ (FN) และจำนวนของสัมประสิทธิ์ของตัวแปรอิสระที่มีค่าไม่เท่ากับ 0 จากการทดสอบสมมติฐาน เมื่อควบคุม FDR ที่ระดับ 0.1 ของข้อมูลจำลองขนาด 100 พบว่า

ตารางที่ 5.1.2 แสดงวิธีการคัดกรองตัวแปรที่เหมาะสมที่สุด เมื่อพิจารณาจากค่า $|\hat{S}|$, FP และ FN ระหว่างวิธี Lasso, Adaptive Lasso, EN และ SCAD จากการวิเคราะห์ขนาดตัวอย่าง (n) เท่ากับ 100 โดยจำแนกตามอัตราส่วน ขนาดตัวอย่างต่อจำนวนตัวแปร (n:p) , ขนาด(Effect size) ของสัมประสิทธิ์ที่ไม่เท่ากับ 0 , จำนวนสัมประสิทธิ์ (β) ที่ไม่เท่ากับ 0 และความสัมพันธ์ (Correlation) ของตัวแปรอิสระ

n : p	|S|	ความสัมพันธ์ (Correlation) ของตัวแปรอิสระ					
		$\rho = 0$	$\rho = 0.5$	$\rho = 0.9$	$\rho = 0$	$\rho = 0.5$	$\rho = 0.9$
		Small Size $(0 < \|\beta\| < 1)$			Large Size $(1 < \|\beta\| < 10)$		
พิจารณาจากค่า $\|\hat{S}\|$							
100:200	10	SC	AL	SC	AL+SC	AL	AL
	20	AL	AL	SC	AL	AL	AL
	50	AL+SC	AL	SC	SC	AL+ SC	SC
100:500	10	AL	SC	AL	SC	SC	SC
	20	SC	SC	AL	AL	AL	AL
	50	SC	AL	SC	AL+ SC	AL	SC
100:1000	10	AL	SC	AL	SC	SC	SC
	20	AL	SC	AL	AL	AL	AL
	50	-	L+AL+SC	AL	-	AL	AL
พิจารณาจากค่า FP							
100:200	10	-	L+EN	L+EN	-	L+EN+SC	L+EN
	20	-	-	L+EN	-	L+AL+EN	L+EN
	50	-	-	L+EN	-	-	L+EN
100:500	10	-	-	EN	-	-	EN
	20	-	-	L+EN	-	-	L+EN
	50	-	-	L+EN	-	-	L+EN
100:1000	10	-	-	EN	-	-	EN
	20	-	-	L+EN	-	-	L+EN
	50	-	-	L+EN	-	EN+SC	-
พิจารณาจากค่า FN							
100:200	10	SC	AL	AL	AL+SC	AL+SC	AL
	20	AL	AL	SC	AL	AL	AL
	50	AL+SC	AL	SC	SC	AL+SC	SC
100:500	10	AL	AL	AL	SC	SC	SC
	20	SC	SC	AL	AL	AL	AL
	50	SC	AL	AL	AL+SC	AL	SC
100:1000	10	AL	SC	AL	SC	SC	SC
	20	AL	AL	AL	AL	AL	AL
	50	-	L+AL+SC	AL	-	AL	AL

หมายเหตุ

L หมายถึง วิธี LASSO

AL หมายถึง วิธี Adaptive Lasso

SC หมายถึง วิธี SCAD

จากตารางที่ 5.2 สามารถสรุปผลได้ว่า เมื่อขนาดตัวอย่าง(n) เท่ากับ 100 พิจารณาจากค่า $|\hat{S}|$,FP และ FN จากค่าของ $|\hat{S}|$ และ FN จะได้วิธีคัดกรองตัวแปรที่เหมาะสมคล้ายกัน นั่นคือวิธี Adaptive Lasso และวิธี SCAD จะมีความเหมาะสมมากที่สุด โดยการคัดกรองตัวแปรในทุกวิธีจะสามารถทำงานได้ดีในกรณีที่จำนวนของสัมประสิทธิ์ของตัวแปรอิสระในตัวแบบที่แท้จริงมีค่าไม่เท่ากับ 0 มีจำนวนน้อยๆ แต่จากค่าของ FP การคัดกรองคัวแปรด้วยวิธี Lasso และวิธี EN ที่เหมาะสมที่สุด นั่นแสดงให้เห็นว่าวิธี Lasso และวิธี EN มีประสิทธิภาพในการคัดเลือกตัวแปรและมีอำนาจในการทดสอบน้อยกว่าวิธี Adaptive Lasso และวิธี SCAD

ส่วนที่ 3 ผลเปรียบเทียบความผิดพลาดในการตรวจจับเชิงบวก (FP) ความผิดพลาดในการตรวจจับเชิงลบ (FN) และจำนวนของสัมประสิทธิ์ของตัวแปรอิสระที่มีค่าไม่เท่ากับ 0 จากการทดสอบสมมติฐาน เมื่อควบคุม FDR ที่ระดับ 0.1 ของข้อมูลจำลองขนาด 200 พบว่า

ตารางที่ 5.1.3 แสดงวิธีการคัดกรองตัวแปรที่เหมาะสมที่สุด เมื่อพิจารณาจากค่า $|\hat{S}|$, FP และ FN ระหว่างวิธี Lasso, Adaptive Lasso, EN และ SCAD จากการวิเคราะห์ขนาดตัวอย่าง (n) เท่ากับ 200 โดยจำแนกตามอัตราส่วน ขนาดตัวอย่างต่อจำนวนตัวแปร (n:p) , ขนาด(Effect size) ของสัมประสิทธิ์ที่ไม่เท่ากับ 0 , จำนวนสัมประสิทธิ์ (β) ที่ไม่เท่ากับ 0 และความสัมพันธ์ (Correlation) ของตัวแปรอิสระ

n : p	$\|S\|$	ความสัมพันธ์ (Correlation) ของตัวแปรอิสระ					
		$\rho = 0$	$\rho = 0.5$	$\rho = 0.9$	$\rho = 0$	$\rho = 0.5$	$\rho = 0.9$
		Small Size $(0 < \|\beta\| < 1)$			Large Size $(1 < \|\beta\| < 10)$		
พิจารณาจากค่า $\|\hat{S}\|$							
200:400	20	SC	SC	SC	L+AL+SC	SC	AL
	40	AL	AL	SC	AL	AL	AL
	100	SC	SC	SC	AL	SC	SC
200:1000	20	SC	SC	SC	SC	SC	SC
	40	AL	AL	AL	AL	AL	AL
	100	L+AL+SC	AL+SC	AL	AL	AL	SC
200:2000	20	SC	AL	AL	SC	SC	SC
	40	SC	AL	AL	AL	AL	AL
	100	-	L+AL+SC	AL	AL+SC	-	SC
พิจารณาจากค่า FP							
200:400	20	-	L+EN	EN	-	L+EN+SC	EN
	40	-	L+EN	L+EN	-	L+EN	EN
	100	-	-	L+EN	-	-	L+EN
200:1000	20	-	L+EN+SC	EN	-	L+EN+SC	L+EN
	40	-	-	L+EN	-	-	-
	100	-	-	L+EN	-	-	L+EN
200:2000	20	-	L+EN	EN	-	-	-
	40	-	-	EN	L+AL+SC	-	L+EN
	100	-	-	EN+SC	-	-	L+AL+EN
พิจารณาจากค่า FN							
200:400	20	SC	SC	AL	L+AL+SC	AL+SC	AL
	40	AL	AL	AL	AL	AL	AL
	100	SC	SC	SC	AL	SC	SC
200:1000	20	SC	SC	AL	SC	SC	SC
	40	AL	AL	AL	AL	AL	AL
	100	L+AL+SC	EN+SC	AL	AL	AL	SC
200:2000	20	SC	AL	AL	SC	SC	SC
	40	SC	AL	SC	AL	AL	AL
	100	-	L+AL+SC	AL+SC	AL+SC	-	-

หมายเหตุ

L หมายถึง วิธี LASSO

AL หมายถึง วิธี Adaptive Lasso

SC หมายถึง วิธี SCAD

จากตารางที่ 5.3 สามารถสรุปผลได้ว่า เมื่อขนาดตัวอย่าง(n) เท่ากับ 200 พิจารณาจากค่า $|\hat{S}|$,FP และ FN ที่ค่าของ $|\hat{S}|$ และ FN จะได้วิธีการคัดกรองตัวแปรที่เหมาะสมคล้ายๆกัน นั่นคือวิธี Adaptive Lasso และวิธี SCAD จะมีความเหมาะสมมาก โดยการคัดกรองตัวแปรในทุกวิธีจะสามารถทำงานได้ดีในกรณีที่จำนวนของสัมประสิทธิ์ของตัวแปรอิสระในตัวแบบที่แท้จริงมีค่าไม่เท่ากับ 0 มีจำนวนน้อยๆ แต่จากค่าของ FP จะได้วิธี Lasso และวิธี EN ที่เหมาะสมที่สุด นั่นแสดงให้เห็นว่าวิธี Lasso และวิธี EN มีประสิทธิภาพในการคัดเลือกตัวแปรและอำนาจในการทดสอบน้อยกว่าวิธี Adaptive Lasso และวิธี SCAD

5.1.2 ผลจากความแตกต่างระหว่างขนาดตัวอย่าง (n)

จากผลที่ได้จะพบว่าที่ n มีขนาดเท่ากับ 10 จะมีความสามารถในการคัดกรองตัวแปรทั้งค่าของ $|\hat{S}|$, FP และ FN จะมีค่าไม่ชัดเจนในทุกกรณี ซึ่งอาจจะเป็นเพราะเมื่อ n = 10 ในขั้นตอนของการทำ Sample Split นั้นจะเหลือ $n_{in} = 5$ ส่งผลให้การคัดกรองตัวแปรในทุกวิธีไม่มีประสิทธิภาพเลย แต่เมื่อ n มีขนาดเท่ากับ 100 และ 200 ค่าของเกณฑ์การพิจารณาข้างต้นจะมีค่าที่ชัดเจน

5.1.3 ผลจากความแตกต่างระหว่างขนาดตัวอย่างต่อจำนวนตัวแปรอิสระ (n:p)

จากผลที่ได้จะพบว่าเมื่ออัตราส่วนระหว่างขนาดตัวอย่างต่อจำนวนตัวแปรอิสระ (n:p) ยิ่งมีค่าห่างกันมาก การคัดกรองตัวแปรของวิธี Lasso, Adaptive Lasso, EN และ SCAD จะมีประสิทธิภาพลดลงในทุกวิธี นั่นแสดงว่าทั้ง 4 วิธีข้างต้นจะมีประสิทธิภาพสูงสุดเมื่อขนาดของ n และ p มีค่าใกล้กันมากที่สุด หรือทั้ง 4 วิธีจะมีประสิทธิภาพลดลงในกรณีที่ข้อมูลมีมิติสูงขึ้น

5.1.4 ผลจากความแตกต่างของจำนวนสัมประสิทธิ์ (β) ของตัวแบบที่แท้จริงไม่เท่ากับศูนย์

จากผลที่ได้จะพบว่าเมื่อจำนวนสัมประสิทธิ์ (β) ของตัวแบบที่แท้จริงไม่เท่ากับศูนย์ที่ร้อยละ 10 ของขนาดตัวอย่างซึ่งเป็นจำนวนที่น้อยที่สุดในแต่ละกรณี การหาค่า $|\hat{S}|$, FP และ FN จะทำได้ดีแต่เมื่อเพิ่มจำนวนสัมประสิทธิ์ (β) ของตัวแบบที่แท้จริงไม่เท่ากับศูนย์เป็นร้อยละ 20 และ 50 ของขนาดตัวอย่าง จะเห็นว่าความสามารถของการหาค่า $|\hat{S}|$, FP และ FN จะลดลงเมื่อร้อยละของจำนวนสัมประสิทธิ์ (β) ของตัวแบบที่แท้จริงไม่เท่ากับศูนย์เมื่อเทียบขนาดตัวอย่างเพิ่มขึ้น เนื่องจากค่าของ $P_j^{(b)}$ เพิ่มขึ้นดังสมการ (2.13) เนื่องจากการคัดกรองตัวแปรทั้ง 4 วิธีข้างต้นเหมาะสำหรับข้อมูลที่มีค่าของสัมประสิทธิ์ส่วนใหญ่เป็น 0

5.1.5 ผลจากความแตกต่างของความสัมพันธ์ (Correlation) ของตัวแปรอิสระ

จากผลที่ได้จะพบว่าเมื่อ Correlation ที่มีค่าเท่ากับ 0 การหาค่า $|\hat{S}|$, FP และ FN จะมีประสิทธิภาพ แต่เมื่อเพิ่มขนาดของ Correlation เป็น 0.5 และ 0.9 จะพบว่าประสิทธิภาพในการหาค่า $|\hat{S}|$, FP และ FN จะลดลงเมื่อค่า Correlation เพิ่มขึ้น

5.1.6 ผลจากความแตกต่างของขนาด (Effect size) ของสัมประสิทธิ์ที่ไม่เท่ากับ 0

จากผลที่ได้จะพบว่าที่ Effect Size มีขนาดใหญ่ $(1 < |\beta| < 10)$ ประสิทธิภาพในการหาค่า $|\hat{S}|$, FP และ FN จะทำได้ดีกว่าเมื่อ Effect Size มีขนาดเล็ก $(0 < |\beta| < 1)$ นั่นแสดงว่าทั้ง 4 วิธี มีประสิทธิภาพในการตรวจจับสัญญาณที่มีขนาดใหญ่ แต่ยังมีข้อจำกัดในการตรวจจับสัญญาณที่มีขนาดเล็ก

5.2 ข้อเสนอแนะ

จากงานวิจัยนี้ผู้ที่สนใจอาจจะนำไปศึกษาต่อได้อีกในเรื่องของ

1. วิธีการคัดกรองตัวแปร ในงานวิจัยนี้เลือกมาศึกษาเพียง 4 วิธีเท่านั้น ในความเป็นจริง
   แล้วยังมีอีกหลายวิธีที่น่าสนใจโดยผู้ที่สนใจอาจจะนำวิธีการคัดกรองตัวแปรอื่นๆ มาร่วมพิจารณาเพื่อเปรียบเทียบประสิทธิภาพได้อีก
2. ขอบเขตในการวิจัย ในเรื่องของขนาดตัวอย่าง, อัตราส่วนระหว่างขนาดตัวอย่างต่อจำนวนตัวแปร (n:p) , ขนาด(Effect size) ของสัมประสิทธิ์ที่ไม่เท่ากับ 0 , จำนวนสัมประสิทธิ์ (β) ที่ไม่เท่ากับ 0 และความสัมพันธ์ (Correlation) ของตัวแปรอิสระ อาจจะมีการเพิ่มหรือลดให้มีความหลากหลายมากยิ่งขึ้นได้
3. กรณีที่ $\mathbf{Y}$ ไม่มีการแจกแจงแบบ Normal
4. กรณีที่รูปแบบความสัมพันธ์ของ $\mathbf{X}$ เปลี่ยนไป
5. ค่าเฉลี่ยของการรันโปรแกรม 100 รอบอาจไม่ใช่ค่าที่ดีที่สุด ดังนั้นการรายงานผลการวิจัยอาจใช้ค่ามัธยฐานแทนหรือค่าเปอร์เซนไทล์ที่ k เมื่อ k มีค่าเล็กเพื่อแสดงถึงค่าที่ดีหรือมีประสิทธิภาพสูงในแต่ละกรณี

รายการอ้างอิง

Benjamini, Y. and Y. Hochberg (1995). "Controlling the False Discovery Rate : A Practical and Powerful Approach to Multiple Testing." Journal of the Royal Statistical Society, Series B **57**(1): 289-300.

Benjamini, Y. and D. Yekutieli (2001). "The Control of the False Discovery Rate in multiple testing under dependency." The Annals of Statistice **29**: 1165-1188.

Fan, J. and R. Li (2001). "Variable selection via nonconcave penalized likelihood and its oracle properties." Journal of the American Statistical Association, Version 3 **96**: 348-1360.

Meinshausen, N., et al. (2009). "P – Values for High – Dimensional Regression." Journal of the American Statistical Association, Version 3: 1–25.

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

Zou, H. (2006). "The Adaptive Lasso and Its Oracle Properties." Journal of American Statistical Association **101**(476): 1418-1429.

Zou, H. and T. Hastie (2003). "Regularization and variable selection via the elastic net." J. R. Statist.Soc. Series B **67**(2): 301-320.

ภาคผนวก

คำสั่งการวิเคราะห์ข้อมูลด้วยโปรแกรม R

ตัวอย่างกรณีที่มีขนาดตัวอย่างเท่ากับ 100 และจำนวนตัวแปรอิสระเท่ากับ 200 เมื่อจำนวนสัมประสิทธิ์ของตัวแบบที่แท้จริงที่ไม่เท่ากับ 0 คิดเป็น 0.1 เท่าของขนาดตัวอย่างที่ระดับความสัมพันธ์ของตัวแปรอิสระเป็น 0 และมีขนาดของสัมประสิทธิ์ที่ไม่เท่ากับ 0 เป็นขนาดเล็ก โดยมีการคัดกรองตัวแปรด้วยวิธี

- Lasso
- Adaptive Lasso
- EN
- SCAD

```
library(mvtnorm)
library(lars)
library(Matrix)
library(parcor)
library(elasticnet)
library(ncvreg)


########################################
########## Case n=100 : p=200 ##########
########################################

n<-100
p<-200

mean_X          <-matrix(c(numeric(p)),nrow=p,ncol=1)
cor_X           <-matrix(c(diag(p)),nrow=p,ncol=p,byrow=TRUE)
rho<-0
cor_X           <- matrix(, nrow = p, ncol = p)

for(s in 1:p){
for(t in 1:p){
        if(t!=s){cor_X[t,s] = rho^abs(t-s)}
        else{cor_X[t,s] = 1}
}}
```

```
mean_e                  <-matrix(c(numeric(p)),nrow=p,ncol=1)
var_e                   <-matrix(c(diag(p)),nrow=p,ncol=p,byrow=TRUE)

for(dd in 1:100){

########################################
########## Simulation Data X ##############
########################################

X<-rmvnorm(n,mean_X,cor_X)

########################################
########## Simulation Error ##############
########################################

error_value     <-array(rmvnorm(n,mean_e,var_e),dim=c(n,1))
error           <-as.matrix(error_value)

########################################
########## Simulation Beta ##############
########################################

nonze_beta                  <- n*0.1
zero_beta                   <- p-nonze_beta
pos_zerobeta                <- sample(1:p, zero_beta, replace=F)
beta_matrix                 <- matrix(NA,p,1)
beta_matrix[pos_zerobeta]   <-0
pos_val                     <- which(is.na(beta_matrix))
beta_matrix[pos_val]        <-runif(nonze_beta,-1,1)
beta                                <-as.matrix(beta_matrix)

write.table(t(beta),paste("D:/lasso/beta.csv"),
      row.names=FALSE,col.name=FALSE,append=T,sep=",")

########################################
### Calculate Y = X(beta) + error ##########
```

```
########################################

Y        <-X%*%beta+error

kk<-1
while(kk<=50){

in.index          <-sample(1:n, round(n/2))
Xin      <-X[in.index,]
Xout              <-X[-in.index,]

Yin      <-as.matrix(Y[in.index,])
Yout              <-as.matrix(Y[-in.index,])

########################################
################ Lasso ##################
########################################

lassomodel     <-
lars(Xin,Yin,type="lasso",use.Gram=FALSE,normalize=TRUE,intercept=T)
cvres             <-cv.lars(Xin,as.numeric(Yin),K=10,type='lasso',plot.it=FALSE)
sAtbest           <-cvres$index[which.min(cvres$cv)]
tmp               <-predict.lars(lassomodel,          type="coefficients",          s=sAtbest,
mode="fraction")
b1                <-as.matrix(tmp$coefficients)
X_zero            <-as.data.frame(t(rep(0,p)))
beta_lasso        <-matrix(rep(0,p),ncol=1)
beta_lasso[1:p,1]<-b1
print(paste("ROUND ",kk))

count.lasso.nonze<-nnzero(beta_lasso, na.counted = NA)

########################################
######### Multi  Sample Split #############
########################################
```

```
Xlout<-Xout[,which(beta_lasso[,1]!=0)]
Xlasso_out<-as.matrix(Xlout,n/2,)

if(count.lasso.nonze!=0){
    lasso_model<-lm(Yout~Xlasso_out)
    sum_lasso<-summary(lasso_model)
    pVal.lasso<-as.matrix(sum_lasso$coef[,4])

beta.lasso          <- matrix(1,p,1)
pos.lasso.pval      <- as.matrix(which(beta_lasso[,1]!=0))

beta.lasso[pos.lasso.pval[,1]]<-pVal.lasso[-1,1]
beta.lasso<-as.matrix(beta.lasso)

    pval.lasso.adj  <- matrix(1,p,1)

pval.lasso.adj<-beta.lasso*count.lasso.nonze
pval.lasso.adj[pval.lasso.adj[,1]>1,1]<-1

    beta.lasso.adj<-t(pval.lasso.adj)
}

if(count.lasso.nonze==0){
    beta.lasso      <- matrix(1,p,1)
    pval.lasso.adj  <- matrix(1,p,1)
    beta.lasso.adj<-t(pval.lasso.adj)

}

########################################
########## Adaptive lasso ###############
########################################

model               <-adalasso(Xin, Yin,k=10, use.Gram=F)
beta_adalasso       <-matrix(rep(0,p), ncol=1)
beta_adalasso[1:p,1] <-model$coefficients.adalasso
```

```
count.alasso.nonze<-nnzero(beta_adalasso, na.counted = NA)

########################################
######### Multi  Sample Split ##########
########################################

Xalout<-Xout[,which(beta_adalasso[,1]!=0)]
Xalasso_out<-as.matrix(Xalout,n/2,)

if(sum(is.na(beta.lasso))==0){
if(count.alasso.nonze!=0){
      alasso_model<-lm(Yout~Xalasso_out)
      sum_alasso<-summary(alasso_model)

      pVal.alasso<-as.matrix(sum_alasso$coef[,4])

beta.alasso                  <- matrix(1,p,1)
pos.alasso.pval              <- as.matrix(which(beta_adalasso[,1]!=0))

beta.alasso[pos.alasso.pval[,1]]<-pVal.alasso[-1,1]
beta.alasso<-as.matrix(beta.alasso)

pval.alasso.adj <- matrix(1,p,1)

pval.alasso.adj<-beta.alasso*count.alasso.nonze
pval.alasso.adj[pval.alasso.adj[,1]>1,1]<-1

      beta.alasso.adj<-t(pval.alasso.adj)
}
if(count.alasso.nonze==0){
      beta.alasso       <- matrix(1,p,1)
      pval.alasso.adj   <- matrix(1,p,1)

      beta.alasso.adj<-t(pval.alasso.adj)
}
```

```
} else {beta.alasso<-NA}

########################################
########## Elastic Net #################
########################################

beta_EN<-matrix(1,p,1)
cv<-cv.glmnet(x=Xin, y=Yin, family="gaussian", nfolds=10)
fit<-glmnet(x=Xin, y=Yin, family="gaussian", alpha=0.5, lambda=cv$lambda.min)
beta_EN<-fit$beta

count.EN.nonze<-nnzero(beta_EN, na.counted = NA)

########################################
######### Multi  Sample Split ##########
########################################

XENout<-Xout[,which(beta_EN[,1]!=0)]
XEN_out<-as.matrix(XENout,n/2,)

if(sum(is.na(beta.lasso))==0 & sum(is.na(beta.alasso))==0){
if(count.EN.nonze!=0){
    EN_model<-lm(Yout~XEN_out)
    sum_EN<-summary(EN_model)

    pVal.EN<-as.matrix(sum_EN$coef[,4])

beta.EN                         <- matrix(1,p,1)
pos.EN.pval             <- as.matrix(which(beta_EN[,1]!=0))

beta.EN[pos.EN.pval[,1]]<-pVal.EN[-1,1]
beta.EN<-as.matrix(beta.EN)

pval.EN.adj     <- matrix(1,p,1)

pval.EN.adj<-beta.EN*count.EN.nonze
```

```
pval.EN.adj[pval.EN.adj[,1]>1,1]<-1

      beta.EN.adj<-t(pval.EN.adj)
}
if(count.EN.nonze==0){
      beta.EN     <- matrix(1,p,1)
      pval.EN.adj          <- matrix(1,p,1)
      beta.EN.adj<-t(pval.EN.adj)

}
} else {beta.EN<-NA}

########################################
############### SCAD #################
########################################

scadcv <-cv.ncvreg(Xin,Yin,nfolds=10,family="gaussian",penalty="SCAD")
optlambda      <-scadcv$lambda[scadcv$min]
model          <-ncvreg(Xin,Yin,family="gaussian",penalty="SCAD")
beta_SCAD1     <-as.matrix(model$beta[,which(model$lambda==optlambda)])
beta_SCAD      <-as.matrix(beta_SCAD1[2:(p+1),1])

count.SCAD.nonze<-nnzero(beta_SCAD, na.counted = NA)

########################################
######### Multi  Sample Split ##########
########################################

Xsout<-Xout[,which(beta_SCAD[,1]!=0)]
XSCAD_out<-as.matrix(Xsout,n/2,)

if(sum(is.na(beta.lasso))==0 & sum(is.na(beta.alasso))==0 & sum(is.na(beta.EN))==0){
if(count.SCAD.nonze!=0){
      SCAD_model<-lm(Yout~XSCAD_out)
      sum_SCAD<-summary(SCAD_model)
```

```
      pVal.SCAD<-as.matrix(sum_SCAD$coef[,4])

beta.SCAD              <- matrix(1,p,1)
pos_pval.SCAD          <- as.matrix(which(beta_SCAD[,1]!=0))

beta.SCAD[pos_pval.SCAD[,1]]<-pVal.SCAD[-1,1]
beta.SCAD<-as.matrix(beta.SCAD)

      pval.SCAD.adj    <- matrix(1,p,1)

pval.SCAD.adj<-beta.SCAD*count.SCAD.nonze
pval.SCAD.adj[pval.SCAD.adj[,1]>1,1]<-1

      beta.SCAD.adj<-t(pval.SCAD.adj)
}

if(count.SCAD.nonze==0){
      beta.SCAD<- matrix(1,p,1)
      pval.SCAD.adj    <- matrix(1,p,1)
      beta.SCAD.adj<-t(pval.SCAD.adj)
}
} else {beta.SCAD<-NA}

if(sum(is.na(beta.lasso))>0 | sum(is.na(beta.alasso))>0 | sum(is.na(beta.EN))>0 |
sum(is.na(beta.SCAD))>0)
{kk<-kk}
if(sum(is.na(beta.lasso))==0 & sum(is.na(beta.alasso))==0 & sum(is.na(beta.EN))==0 &
sum(is.na(beta.SCAD))==0)
{
write.table(beta.lasso.adj,paste("D:/lasso/pVal_Lasso",dd,".csv"),
      row.names=FALSE,col.name=FALSE,append=T,sep=",")

write.table(beta.alasso.adj,paste("D:/alasso/pVal_adaLasso",dd,".csv"),
      row.names=FALSE,col.name=FALSE,append=T,sep=",")

write.table(beta.EN.adj,paste("D:/EN/pVal_EN",dd,".csv"),
```

```
      row.names=FALSE,col.name=FALSE,append=T,sep=",")

write.table(beta.SCAD.adj,paste("D:/SCAD/pVal_SCAD",dd,".csv"),
      row.names=FALSE,col.name=FALSE,append=T,sep=",")
      kk<-kk+1
}}}
```

ประวัติผู้เขียนวิทยานิพนธ์

นายศุภวัฒน์ อังคะสี เกิดวันอังคารที่ 22 พฤษภาคม พ.ศ. 2527 สำเร็จการศึกษาปริญญาวิทยาศาสตรบัณฑิต (วท.บ.) สาขาวิชาสถิติ ภาควิชาคณิตศาสตร์และสถิติ คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยธรรมศาสตร์ ในปีการศึกษา 2549 และเข้าศึกษาต่อในหลักสูตรวิทยาศาสตรมหาบัณฑิต (วท.ม.) สาขาวิชาสถิติ ภาควิชาสถิติ คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย ในปีการศึกษา 2555