ข้อเสนอโครงงานวิศวกรรมไฟฟ้า วิชา 2102490 ปีการศึกษา 2558

ชื่อโครงงาน (ไทย) การหาโครงสร้างเชิงสาเหตุสำหรับข้อมูลอนุกรมเวลาด้วยวิธีทดสอบทางสถิติบนการประมาณแบบจำลอง

Project Name (English) Exploring Granger causality for time series via statistical tests on estimated models

ชื่อนิสิต นาย นันทนัช รักษาศรี **หมายเลขประจำตัว** 5530299121 **SRA** ACO

อาจารย์ที่ปรึกษา อาจารย์ จิตโกมุท ส่งศิริ **ห้องปฏิบัติการวิจัย** ระบบควบคุม

1 หลักการและเหตุผล

ในปัจจุบัน ข้อมูลอนุกรมเวลา (time series) นั้นเกี่ยวข้องกับชีวิตประจำวันไม่มากก็น้อย ตัวอย่างเช่น ราคาของหุ้นในแต่ละวัน ปริมาณน้ำฝนในแต่ละเดือน ปริมาณความต้องการใช้ไฟฟ้าในแต่ละวัน หรือแม้กระทั่งสัญญาณคลื่นไฟฟ้าทางสมอง ถ้าสามารถทราบถึงความสัมพันธ์ระหว่างตัวแปรต่างๆของข้อมูลอนุกรมเวลา ก็จะสามารถอธิบายการเปลี่ยนแปลงหรือพลวัตของข้อมูลอนุกรมเวลาว่ามีผลจากอะไรบ้าง ข้อมูลอนุกรมเวลาที่สนใจในโครงงานนี้คือ EEG (electroencephalogram) หรือการตรวจคลื่นไฟฟ้าทางสมอง เป็นการบันทึกสัญญาณไฟฟ้าที่เกิดจากผลรวมของกระแสไฟฟ้าของกลุ่มเซลล์ในสมอง โดยเลือกใช้แบบจำลองถดถอยตัวเอง (Autoregressive (AR) model) มาอธิบาย เนื่องจากแบบจำลอง AR นั้นค่าสัญญาณขาออกที่เวลาปัจจุบันจะขึ้นกับค่าของสัญญาณขาออกของเวลาในอดีตและสัญญาณรบกวนในปัจจุบันเท่านั้น ดังสมการ

$$y(t) = c + A_1 y(t-1) + A_2 y(t-2) + \cdots + A_p y(t-p) + v(t) \tag{1}$$

เมื่อ $y(t)$ และ $v(t)$ คือสัญญาณขาออก และสัญญาณรบกวนที่เวลา t ตามลำดับ ดังนั้นแบบจำลอง AR จึงเป็นแบบจำลองที่ดีแบบจำลองหนึ่งสำหรับใช้ในการอธิบายสัญญาณ EEG

สำหรับโครงงานนี้ การอธิบายความสัมพันธ์ของตัวแปรในข้อมูลอนุกรมเวลา หรือโครงสร้างเชิงสาเหตุจะถูกอธิบายด้วยเงื่อนไขความสัมพันธ์ระหว่างเหตุและผล (Granger causality) โดยจาก [1] หลักการพื้นฐานของเงื่อนไขความสัมพันธ์ระหว่างเหตุและผลคือ ถ้าหากตัวแปร x มีผลต่อตัวแปร y แล้วการที่รู้ค่าของตัวแปร x ก่อน ย่อมจะช่วยให้การทำนายค่าของตัวแปร y มีความแม่นยำมากขึ้นกว่าการที่ไม่รู้ค่าของตัวแปร x มาก่อน ดังนั้นเมื่อนำมาประยุกต์ใช้กับแบบจำลอง AR เงื่อนไขความสัมพันธ์ระหว่าง

เหตุและผลของแบบจำลองนั้นสามารถเขียนได้ในรูปสมการเชิงเส้นของค่าสัมประสิทธิ์หรือค่าพารามิเตอร์ของแบบจำลอง หรือก็คือ y_j ไม่ใช่สาเหตุแบบ Granger ต่อ y_i ก็ต่อเมื่อ

$$(A_k)_{ij} = 0 \tag{2}$$

เมื่อ $(A_k)_{ij}$ คือสมาชิกตัวที่ (i, j) ของเมทริกซ์ A_k ซึ่งคือพารามิเตอร์ของแบบจำลอง AR สำหรับทุกค่า $k = 1, 2, ..., p$

การประมาณค่าพารามิเตอร์ของแบบจำลองจะเลือกใช้วิธีการประมาณแบบความควรจะเป็นสูงสุด (maximum likelihood (ML) estimation) ถ้าค่าของพารามิเตอร์มีค่าเท่ากับศูนย์ จะสามารถบอกได้ว่าข้อมูลอนุกรมเวลาไม่สัมพันธ์กัน โดยปรกติแล้วเมทริกซ์ A_k ที่ประมาณได้อาจจะมีบางสมาชิกที่ไม่เท่ากับศูนย์ จึงต้องตั้งสมมติฐานขึ้นมาเพื่อทดสอบสมการ (2) ว่าเป็นจริงหรือไม่ กล่าวคือการทดสอบหารูปแบบโครงสร้างของค่าที่เป็นศูนย์ในพารามิเตอร์ของแบบจำลอง AR

การทดสอบสมมติฐานนั้นจะนำหลักการของการทดสอบ Wald มาใช้ โดยจาก [2] การทดสอบ Wald มีหลักการคือ ถ้าสมมติฐานว่าค่าที่แท้จริงของพารามิเตอร์มีค่าเท่ากับศูนย์เป็นจริง แล้วค่าพารามิเตอร์ที่ประมาณได้ควรมีค่าเข้าใกล้ศูนย์อย่างมีนัยยะสำคัญ จากนั้นจะสามารถประมาณค่าพารามิเตอร์ใหม่ได้ภายใต้ข้อกำหนดที่ว่าค่าของสมาชิกบางตัวของพารามิเตอร์มีค่าเป็นศูนย์ ซึ่งอยู่ในรูปแบบปัญหาการหาค่าที่เหมาะสมที่สุดภายใต้ข้อกำหนดค่าคงที่เป็นศูนย์ (optimization problem with zero constrain)

โครงงานนี้มีจุดประสงค์ที่จะอธิบายความสัมพันธ์ระหว่างตัวแปรต่างๆ ของข้อมูลอนุกรมเวลาผ่านค่าพารามิเตอร์ของแบบจำลอง AR โดยการหารูปแบบโครงสร้างของศูนย์ในพารามิเตอร์ของแบบจำลอง

2 การกำหนดปัญหา

พิจารณา (1) แบบจำลองถดถอยตัวเอง (Autoregressive (AR) model)

$$y(t) = c + A_1 y(t-1) + A_2 y(t-2) + \cdots + A_p y(t-p) + v(t)$$

เมื่อ $y(t) \in \mathbb{R}^n = (y_1(t), y_2(t), ..., y_n(t))$ เป็นค่าของสัญญาณขาออกของแบบจำลอง และ $y_i(t)$ เป็นสมาชิกตัวที่ i ของ $y(t)$, $A_1, A_2, ..., A_p \in \mathbb{R}^{n \times n}$ เป็นค่าสัมประสิทธิ์หรือพารามิเตอร์ของแบบจำลอง, $c \in \mathbb{R}^n$ เป็นเวคเตอร์ค่าคงที่, p คืออันดับของแบบจำลองและ $v(t)$ คือสัญญาณรบกวนซึ่งกำหนดให้เป็นสัญญาณรบกวนเกาส์สีขาว (Gaussian white noise) โดยมีค่าเฉลี่ยเป็นศูนย์ และเมทริกซ์ความแปรปรวนเป็น Σ เราสามารถเขียนแบบจำลอง (1) ในรูปแบบสมการเชิงเส้นในพารามิเตอร์ได้ดังนี้

$$y(t) = AH(t) + v(t) \tag{3}$$

เมื่อ $H(t) = \begin{bmatrix} 1 & y(t-1)^T & y(t-2)^T & \cdots & y(t-p)^T \end{bmatrix}^T \in \mathbb{R}^{np+1}$ และ $v(t) \sim \mathcal{N}(0, \Sigma)$ ค่าพารามิเตอร์ของแบบจำลองที่ต้องการประมาณคือ $A = \begin{bmatrix} c & A_1 & A_2 & \cdots & A_p \end{bmatrix} \in \mathbb{R}^{n \times (np+1)}$ ซึ่งจะนำวิธีการประมาณแบบ ML มาใช้ประมาณค่าพารามิเตอร์นั้น โดยจะถูกกล่าวถึงในภาคผนวกหัวข้อที่ 6.1

การทดสอบ Wald

ในหัวข้อนี้ จะกล่าวถึงการทดสอบ Wald ซึ่งเป็นการทดสอบสมมติฐาน (hypothesis test) บนพารามิเตอร์ที่ได้จากการประมาณว่าเป็นไปตามสมมติฐานที่เราตั้งไว้หรือไม่ โดยเป็นเนื้อหาสรุปมาจาก [2] กำหนดให้ตัวแปรไม่ทราบค่า $\theta \in \mathbb{R}^p$ เป็นพารามิเตอร์ที่ถูกประมาณค่าด้วยวิธีการของ ML และให้ $\mathbf{Avar}(\theta)$ คือเมทริกซ์ค่าความแปรปรวนร่วมเชิงเส้นกำกับ (Asymptotic covariance matrix) ของ θ ซึ่งสามารถหาได้จาก

$$\mathbf{Avar}(\theta) = \frac{1}{N}\mathcal{I}(\theta)^{-1}$$

เมื่อ $\mathcal{I}(\theta)$ คือเมทริกซ์ของฟิชเชอร์อินฟอร์เมชัน (Fisher information matrix) ของ θ ซึ่งอยู่ใน y (สำหรับ 1 ตัวอย่าง) โดย y คือตัวอย่างของข้อมูลและ N คือ จำนวนตัวอย่าง โดยนิยามคือ

$$\begin{aligned}\mathcal{I}(\theta) &= \mathbb{E}\left[(\nabla_\theta \log f(y|\theta))(\nabla_\theta \log f(y|\theta))^T\right] \\ &= -\mathbb{E}\left[\nabla_\theta^2 \log f(y|\theta)\right]\end{aligned}$$

จากนั้นตั้งสมมติฐานว่าง (null hypothesis) เพื่อใช้ในการทดสอบสมมติฐานคือ

$$H_0 : r(\theta) = 0$$

เมื่อ $r : \mathbb{R}^p \to \mathbb{R}^m$ คือฟังก์ชันข้อกำหนด (restriction function) โดย $m \leqslant p$ ตัวอย่างเช่น $\theta_1 = 0, \theta_1 + \theta_3 = 0$

และจาก [**??**] การทดสอบ Wald จะใช้การทดสอบเชิงสถิติดังนี้

$$W = r(\hat{\theta})^T \left[D_r(\hat{\theta})\widehat{\mathbf{Avar}}(\hat{\theta})D_r(\hat{\theta})^T\right]^{-1} r(\hat{\theta}) \tag{4}$$

เมื่อ $D_r(\theta)$ คือจาโคเบียนของฟังก์ชัน r เทียบกับ θ และ $\widehat{\mathbf{Avar}}(\hat{\theta})$ คือค่าประมาณของเมทริกซ์ค่าความแปรปรวนช่วงเชิงเส้นกำกับของ $\hat{\theta}$ โดยสำหรับการประมาณแบบ ML ค่าของ $\widehat{\mathbf{Avar}}(\hat{\theta})$ สามารถหาได้จาก

$$\widehat{\mathbf{Avar}}(\hat{\theta}_{ml}) = \left[-\sum_{i=1}^{N} \nabla^2 \log f(y_i|\hat{\theta})\right]^{-1} \tag{5}$$

เราสามารถพิสูจน์ได้ว่าค่าทางสถิติของ Wald (W) จะมีการแจกแจงลู่เข้าสู่การแจกแจงแบบไคสแควร์ (Chi-square distribution) โดยมีองศาอิสระเท่ากับ m

กำหนดให้ α คือค่าระดับนัยสำคัญ (significance level) ซึ่งนิยามคือ

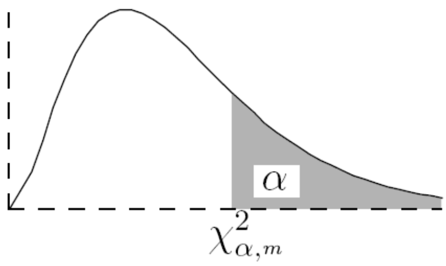
$$\alpha = \mathbf{Prob}(W > \mathcal{X}^2_{\alpha,m}) = 1 - P(W \leqslant \mathcal{X}^2_{\alpha,m}) \approx 1 - F(\mathcal{X}^2_{\alpha,m})$$

โดย $F(\mathcal{X}^2_{\alpha,m})$ คือฟังก์ชันการแจกแจงแบบไคสแควร์ของ $\mathcal{X}^2_{\alpha,m}$ และ $\mathcal{X}^2_{\alpha,m}$ คือค่าวิกฤต (critical value) ดังนั้นจะได้ว่า

$$\mathcal{X}^2_{\alpha,m} = F^{-1}(1-\alpha)$$

โดยการทดสอบสมมติฐานด้วยวิธีการทดสอบ Wald เราจะปฏิเสธ H_0 ถ้า

$$W > \mathcal{X}^2_{\alpha,m}$$



รูป 1: รูปแสดงการแจกแจงไคสแควร์

วัตถุประสงค์ของโครงงงาน

1. นำเสนอรูปแบบปัญหาการเรียนรู้เงื่อนไขความสัมพันธ์ระหว่างเหตุและผล จากแบบจำลอง AR ด้วยวิธีการทดสอบสมมติฐานแบบ Wald
2. พัฒนาการเขียนโปรแกรม Matlab เพื่อแก้ปัญหาการเรียนรู้เงื่อนไขความสัมพันธ์ระหว่างเหตุและผล ดังข้อที่ 1

ขอบเขตของโครงงาน

1. การทดสอบประสิทธิภาพของการทดสอบ Wald จะถูกทำบนข้อมูลที่สังเคราะห์ขึ้นมาเอง และจะนำการทดลองไปใช้จริงบนข้อมูล EEG โดยเป็นส่วนหนึ่งของโครงงานของนาย ปวริศร ใจทหาร
2. ขนาดของปัญหาที่พิจารณาจะอยู่ในอันดับ $n \simeq 20-50$ โดยที่เมื่อปัญหามีอันดับที่สูงขึ้น จะพัฒนาโปรแกรมให้มีประสิทธิภาพมากขึ้นในแง่ของการคำนวณ หรือทางเลือกของภาษา

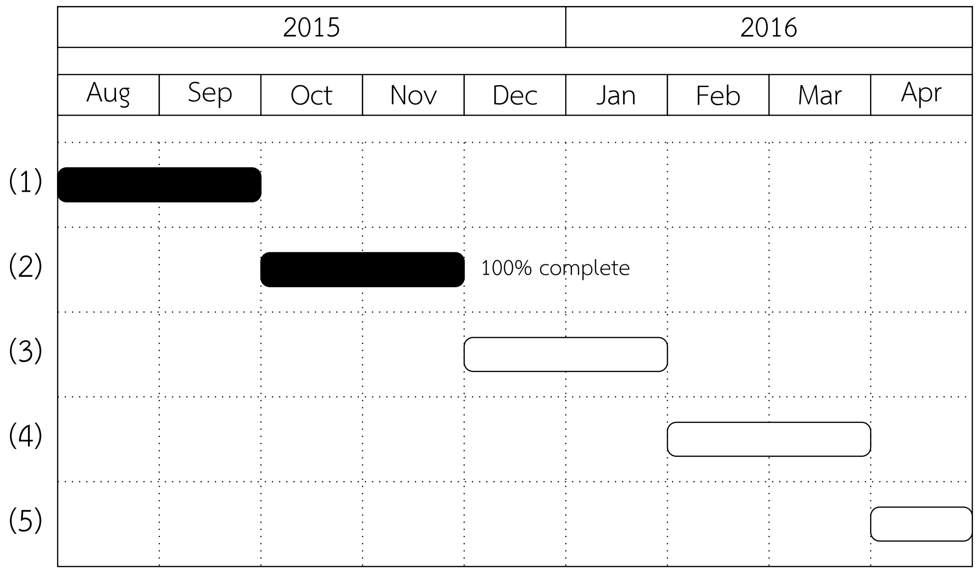
3 ระเบียบวิธีที่ใช้ในโครงงาน

3.1 แนวทางในการทำโครงงาน

1. ศึกษาวิธีการประมาณแบบจำลองด้วยวิธี Least-Square (LS) และ Maximum Likelihood (ML)
2. ศึกษาแบบจำลอง AR และหาค่าพารามิเตอร์ของแบบจำลองด้วยวิธี ML
3. ศึกษาหลักการของเงื่อนไขความสัมพันธ์ระหว่างเหตุและผลกับแบบจำลอง AR
4. ศึกษาการประมาณพารามิเตอร์ของแบบจำลองด้วยวิธี LS ภายใต้เงื่อนไขเชิงเส้น
5. ศึกษาวิธีการทดสอบสมมติฐานด้วยวิธีการทดสอบ Wald และทดลองการทดสอบสมมติฐานกับปัญหาเชิงเส้น และตัวแปรของแบบจำลอง AR
6. ทดลองประมาณค่าพารามิเตอร์ของแบบจำลอง AR เมื่อนำหลักการของเงื่อนไขความสัมพันธ์ระหว่างเหตุและผล และเสถียรภาพมาใช้

7. ทดลองใช้ค่าระดับนัยสำคัญหลายๆค่าเพื่อหาแบบจำลองที่เหมาะสมที่สุด
8. ทดลองทำการทดสอบอัตราส่วนความควรจะเป็น (likelihood ratio test) เพื่อใช้แยกแยะรูปแบบโครงสร้างของเงื่อนไขความสัมพันธ์ระหว่างเหตุและผล
9. เขียนรายงาน

3.2 แผนงาน



รูป 2: Gantt Chart การทำโครงงานวิศวกรรม ปีการศึกษา 2558 จนถึงเดือน เม.ย. 59 โดยมีรายละเอียดดังนี้
(1) ศึกษาและทบทวนวรรณกรรม (LS,MS estimation, Granger causality, AR model)
(2) ศึกษาการประมาณแบบจำลองด้วยวิธี LS ภายใต้เงื่อนไขเชิงเส้น, การทดสอบ Waldและทดลองกับปัญหาเชิงเส้นและพารามิเตอร์ของแบบจำลอง AR
(3) ทดลองประมาณค่าพารามิเตอร์ของแบบจำลอง AR เมื่อนำหลักการของ Granger causality และ Stability มาใช้
(4) ทดลองใช้ค่าระดับนัยสำคัญหลายๆค่าเพื่อหาแบบจำลองที่เหมาะสมที่สุดและทำการทดสอบอัตราส่วนความควรจะเป็นเพื่อใช้แยกแยะรูปแบบของเงื่อนไขความสัมพันธ์ระหว่างเหตุและผล และทดลองกับข้อมูลจริง
(5) เขียนรายงานวิชา 2102499

3.3 ผลลัพธ์ในแต่ละขั้นตอน (Deliveries)

1. สามารถประมาณค่าพารามิเตอร์ของแบบจำลองด้วยวิธี LS และ ML ได้
2. เข้าใจรูปแบบและการนำไปใช้ของแบบจำลอง AR และสามารถประมาณค่าพารามิเตอร์ในแบบจำลองด้วยวิธี ML
3. เข้าใจหลักการของเงื่อนไขความสัมพันธ์ระหว่างเหตุและผล เมื่อนำมาใช้กับแบบจำลอง AR
4. เข้าใจวิธีการแก้ปัญหาของการประมาณพารามิเตอร์ของแบบจำลองภายใต้เงื่อนไขเชิงเส้น
5. สามารถหาค่าประมาณของพารามิเตอร์ของปัญหาเชิงเส้นและของแบบจำลอง AR เมื่อมีข้อจำกัดว่าตำแหน่งใดๆมีค่าเท่ากับศูนย์

6. สามารถประมาณค่าพารามิเตอร์ของแบบจำลองAR ใหม่ โดยสอดคล้องกับหลักการของเงื่อนไขความสัมพันธ์ระหว่างเหตุและผล และแบบจำลองที่ประมาณได้มีเสถียรภาพ
7. สามารถหาแบบจำลองที่เหมาะสมที่สุดที่ใช้ในการประมาณข้อมูลอนุกรมเวลา
8. สามาถแยกแยะรูปแบบโครงสร้างของเงื่อนไขความสัมพันธ์ระหว่างเหตุและผล เพื่อนำมาพิจารณาข้อมูลอนุกรมเวลาได้ว่าควรอยู่ในหมวดหมู่อะไร
9. รายงานฉบับสมบูรณ์

4 ความก้าวหน้าตามแผนงาน

4.1 การทดสอบ Wald กับปัญหากำลังสองน้อยสุด (Least-Square)

การทดลองในส่วนนี้จะถูกกล่าวถึงในภาคผนวกหัวข้อที่ 6.2 โดยจากผลจากผลการทดลองจะเห็นว่าค่าความผิดพลาดแบบที่สองจะมีค่าลดลงเมื่อเพิ่มค่าของระดับนัยสำคัญและจำนวนตัวอย่าง แต่จะมีค่าเพิ่มมากขึ้นเมื่อค่าความแปรปรวนของสัญญาณรบกวนมีค่ามากขึ้น ในขณะที่ค่าความผิดพลาดแบบที่หนึ่งจะมีค่าประมาณเท่ากับค่าระดับนัยสำคัญเสมอ เนื่องจากการทดสอบสมมติฐานได้นิยามให้ค่าระดับนัยสำคัญและค่าความผิดพลาดแบบที่หนึ่งเป็นค่าเดียวกัน

4.2 การทดสอบ Wald กับปัญหากำลังสองน้อยสุด (LS) บนแบบจำลอง AR

เพื่อสังเกตรูปแบบโครงสร้างของศูนย์ในพารามิเตอร์ของแบบจำลองที่ถูกประมาณได้เมื่อใช้ค่าระดับนัยสำคัญต่างๆกัน เริ่มต้นจากการสร้างแบบจำลองตามสมการ (3) โดยใช้ค่า $y(t) \in \mathbb{R}^n$ อันดับของแบบจำลอง $p = 4$ และจำนวนข้อมูลทั้งหมด 1,000 จุด $(t = 1, 2, ..., 1000)$ พารามิเตอร์ของแบบจำลอง $A_k \in \mathbb{R}^{n \times n}$ สำหรับ $k = 1, 2, 3, 4$ จะถูกสร้างให้มีบางสมาชิก $(A_k)_{ij}$ เป็นศูนย์จำนวนครึ่งหนึ่งของจำนวนพารามิเตอร์ทั้งหมด หรือก็คือ 50% และ $v(t) \sim \mathcal{N}(0, 10I_n)$ โดยใช้ค่า $n = 20$ แล้วจึงทำการประมาณค่าพารามิเตอร์ A, Σ ซึ่งคือผลเฉลยแบบปิด (closed-form) ของปัญหา ML จากสมการ (10), (11) ในส่วนของภาคผนวกหัวข้อที่ 6.1 ตามลำดับ

จากนั้นจึงใช้วิธีการจากภาคผนวกหัวข้อที่ 6.3 เพื่อนำพารามิเตอร์ของแบบจำลองไปทดสอบ Wald ซึ่งจะได้รูปแบบของสมการใหม่ในรูปของสมการ (14) และฟังก์ชันข้อกำหนดสำหรับทดสอบคู่ (i, j) ใดๆ ดัง (15) ซึ่งสามารถเขียนได้เป็น

$$r(\hat{\theta}) = \left[e_i^T \otimes I_p\right] \hat{\theta}$$

เมื่อ $\otimes$ คือตัวดำเนินการ kronecker product และ $\hat{\theta} \in \mathbb{R}^{n^2 p}$ คือค่าประมาณของ θ ในสมการ (13) ซึ่งมีจำนวนตัวแปรทั้งหมด $n^2 p = 1600$ และค่าประมาณของเมทริกซ์ค่าความแปรปรวนร่วมของ $\hat{\theta}$ ซึ่ง

หาได้จาก (5) คือ

$$\widehat{\mathbf{Avar}}(\hat{\theta}) = \left[\sum_{t=p+1}^{N} \bar{H}(t)^T \hat{\Sigma}^{-1} \bar{H}(t) \right]^{-1}$$

และค่าทางสถิติของ Wald (W) สำหรับคู่ $(\hat{A}_k)_{ij}$ ใดๆจาก (4) จะลดรูปเป็น

$$W_{ij} = \hat{B}_{ij}^T \left[\widehat{\mathbf{Avar}}(\hat{\theta})_{ij} \right]^{-1} \hat{B}_{ij}$$

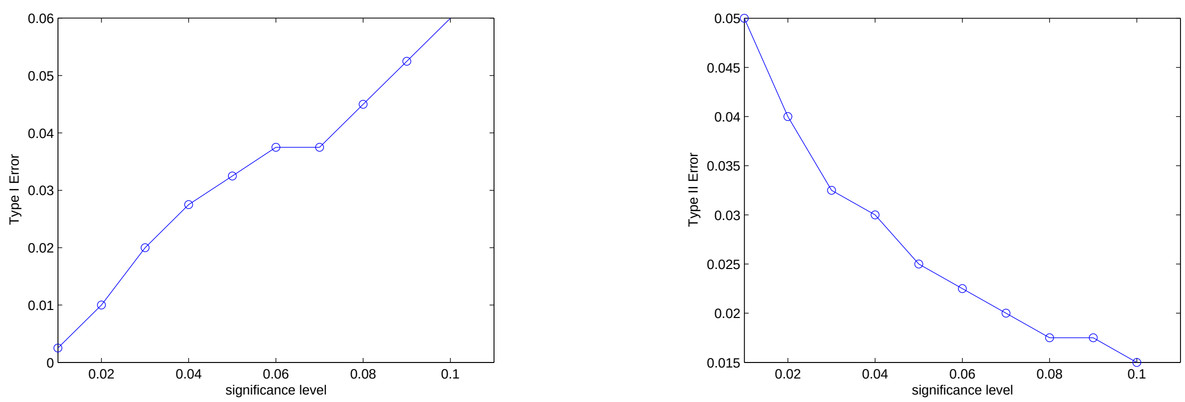
โดยกำหนดให้ $\widehat{\mathbf{Avar}}(\hat{\theta})_{ij}$ คือเมทริกซ์ย่อยขนาด $p \times p$ ในแนวทแยงมุม (main diagonal block) ของ $\widehat{\mathbf{Avar}}(\hat{\theta})$ ที่สอดคล้องกับพารามิเตอร์ B_{ij}

จากนั้นจึงทำการทดสอบ Wald สำหรับ $(A_k)_{ij}$ แต่ละคู่ (i, j) โดยใช้ค่าระดับนัยสำคัญต่างๆกัน และพิจารณาค่าความผิดพลาดทั้งสองแบบของการทดสอบสมมติฐาน ซึ่งคือค่าความผิดพลาดแบบที่หนึ่ง (Type I error) คือค่าความน่าจะเป็นที่ค่าพารามิเตอร์ที่*ประมาณ*ได้ไม่เท่ากับศูนย์เมื่อกำหนดให้*ค่า*จริงของพารามิเตอร์เท่ากับศูนย์

$$\text{Type I Error} = \mathbf{Prob}((\hat{A}_k)_{ij} \neq 0 |\ (A_k)_{ij} = 0)$$

และค่าความผิดพลาดแบบที่สอง (Type II error)คือค่าความน่าจะเป็นที่ค่าพารามิเตอร์ที่*ประมาณ*ได้เท่ากับศูนย์เมื่อกำหนดให้*ค่า*จริงของพารามิเตอร์ไม่เท่ากับศูนย์

$$\text{Type II Error} = \mathbf{Prob}((\hat{A}_k)_{ij} = 0 |\ (A_k)_{ij} \neq 0)$$

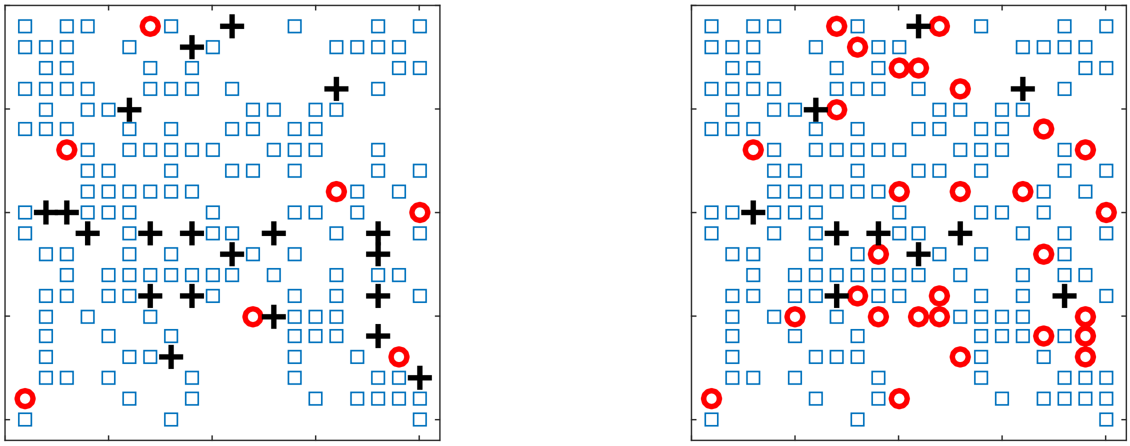


(a) ค่าความผิดพลาดแบบที่หนึ่ง

(b) ค่าความผิดพลาดแบบที่สอง

รูป 3: รูปแสดงค่าความผิดทั้งสองแบบ เมื่อเปลี่ยนแปลงค่าระดับนัยสำคัญ (α)

จากการทดลองจะเห็นว่าค่าความผิดพลาดแบบที่สองจะมีค่าลดลงตามค่าระดับนัยสำคัญที่เพิ่มขึ้น ในขณะที่ค่าความผิดพลาดแบบที่หนึ่งจะมีค่าเพิ่มขึ้นตามค่าระดับนัยสำคัญที่เพิ่มขึ้น ดังรูปที่ 3 และแสดงตัวอย่างรูปแบบโครงสร้างของศูนย์ของพารามิเตอร์ที่ประมาณได้เทียบกับค่าจริงในรูปที่ 4



(a) $\alpha = 0.01$ (b) $\alpha = 0.1$

รูป 4: รูปแสดงตัวอย่างรูปแบบโครงสร้างของศูนย์ของพารามิเตอร์ที่ประมาณได้เมื่อเทียบกับค่าจริง เมื่อ □ แทนจุดข้อมูลที่ไม่เท่ากับศูนย์ทั้งคู่ของค่าจริงและค่าประมาณ ○ แทนจุดที่ค่าจริงเท่ากับศูนย์แต่ค่าประมาณไม่เป็นศูนย์ +แทนจุดที่ค่าจริงไม่เท่ากับศูนย์แต่ค่าประมาณเป็นศูนย์และช่องว่างคือจุดข้อมูลที่เป็นศูนย์ทั้งคู่ของค่าจริงและค่าประมาณ

5 บทสรุป

โครงงานนี้จะทำการหารูปแบบโครงสร้างของศูนย์ของพารามิเตอร์ของแบบจำลอง AR มีจุดประสงค์เพื่ออธิบายความสัมพันธ์ระหว่างตัวแปรต่างๆ ของข้อมูลอนุกรมเวลาผ่านค่าพารามิเตอร์ของแบบจำลอง โดยในส่วนของภาคการศึกษาต้นจะทำการศึกษาวิธีการประมาณแบบกำลังสองน้อยสุดและแบบความควรจะเป็นสูงสุด และศึกษาการทดสอบ Wald เพื่อนำมาใช้หารูปแบบโครงสร้างของศูนย์ของพารามิเตอร์ของแบบจำลอง และภาคการศึกษาปลายจะหาแบบจำลองที่เหมาะสมสำหรับแบบจำลอง เนื่องจากการทดสอบ Wald จะมีผลลัพธ์ของโครงสร้างของศูนย์มากกว่าหนึ่งแบบทำให้ต้องพิจารณาววิธีการเลือกแบบจำลอง และพัฒนาประสิทธิภาพของการเขียนโปรแกรม Matlab สำหรับกรณีที่อันดับของปัญหามีค่าที่สูงขึ้น

References

[1] J. Songsiri, *Sparse Autoregressive Model Estimation for Learning Granger Causality in Time Series*, Proc. of the 38th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Canada, May 2013.

[2] W. H. Greene, *Econometric Analysis*. Prentice Hall, 2008

6 ภาคผนวก

6.1 การประมาณแบบความควรจะเป็นสูงสุด (Maximum Likelihood (ML) Estimation)

พิจารณา $y = (y_1, ..., y_N)$ เป็นตัวอย่างหรือสัญญาณขาออกที่สังเกตได้จากแบบจำลอง, θ คือค่าพารามิเตอร์ไม่ทราบค่าของแบบจำลองที่ต้องการจะประมาณ การประมาณค่า θ จาก y ด้วยหลักการของ ML ทำได้โดยการเลือก θ ที่ ทำให้ค่าฟังก์ชันความหนาแน่นของความน่าจะเป็นมีค่าสูงสุด

$$\hat{\theta} = \arg\max_{\theta} f(y|\theta) \tag{6}$$

โดย $f(y|\theta)$ คือฟังก์ชันความหนาแน่นของความน่าจะเป็นแบบมีเงื่อนไข (conditional likelihood function) ของ y สำหรับค่า θ หนึ่งๆ และนำหลักการของ ML มาประยุกต์ใช้กับแบบจำลอง AR จะได้ว่าฟังชันความควรจะเป็นของแบบจำลองคือ

$$\begin{aligned} f(y|\theta) = f(y|A,\Sigma) &= f(y(p+1), y(p+2), ..., y(N)|y(1), y(2), ..., y(p), A, \Sigma) \\ &= \frac{1}{(2\pi)^{\frac{(N-p)}{2}}} \cdot \frac{1}{\det \Sigma^{\frac{(N-p)}{2}}} \cdot \exp -\frac{1}{2}\sum_{t=p+1}^{N}(y(t) - AH(t))^T\Sigma^{-1}(y(t) - AH(t)) \end{aligned} \tag{7}$$

เนื่องจาก (7) อยู่ในรูปของฟังก์ชันเอกซ์โพเนนเชียล (exponential function) เรามักจะพิจารณาลอการิทึมของ (7) แทน เพื่อใช้แก้ปัญหา (6) ได้สะดวกมากขึ้น

$$\log(f(y|A,\Sigma)) = \mathcal{L}(A,\Sigma) = \frac{N-p}{2}\log\det\Sigma^{-1} - \frac{1}{2}\sum_{t=p+1}^{N}(y(t) - AH(t))^T\Sigma^{-1}(y(t) - AH(t)) \tag{8}$$

เทอมสุดท้ายใน (8) สามารถเขียนใหม่ได้เป็นดังนี้

$$\mathcal{L}(A,\Sigma) = \frac{N-p}{2}\log\det\Sigma^{-1} - \frac{1}{2}\|L[Y - AH]\|_F^2 \tag{9}$$

โดยให้ $L^TL = \Sigma^{-1}$

$$Y = \begin{bmatrix} y(p+1) & y(p+2) & \cdots & y(N) \end{bmatrix}_{n\times(N-p)}, H = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ y(p) & y(p+1) & \cdots & y(N-1) \\ y(p-1) & y(p) & \cdots & y(N-2) \\ \vdots & \vdots & & \vdots \\ y(1) & y(2) & \cdots & y(N-p) \end{bmatrix}_{(np+1)\times(N-p)}$$

จากนั้นสามารถหาค่าประมาณของ A และ Σ ได้โดยจาก (9) จะเห็นว่าปัญหา (6) จะเทียบเท่ากับปัญหากำลังสองน้อยสุด (least-square) ซึ่งจะเป็นปัญหาการหาค่าที่เหมาะสมที่สุดแบบทั่วไป (unconstrain optimization) จะได้ว่าผลเฉลยต้องสอดคล้องกับเงื่อนไขเท่ากับศูนย์ของเกรเดียนต์ของลอการิทึมของฟังก์ชันความน่าจะเป็นแบบมีเงื่อนไขของ y เทียบกับ θ คือ

$$\nabla_\theta \log f(y|\theta) = 0$$

ดังนั้น สำหรับการหาค่าสูงสุดของ $\mathcal{L}(A,\Sigma)$ จึงต้องหา A และ Σ ที่ทำให้เกรเดียนต์ของลอการิทึมของฟังก์ชันความน่าจะเป็นแบบมีเงื่อนไขของ y เทียบกับ A และ Σ^{-1} เท่ากับศูนย์ตามลำดับ

$$\nabla_A \mathcal{L}(A,\Sigma) = 0 \text{ และ } \nabla_{\Sigma^{-1}} \mathcal{L}(A,\Sigma) = 0$$

จากเงื่อนไขข้างต้น เราจะได้ผลเฉลย A, Σ ในรูปแบบปิด (closed-form) คือ

$$\hat{A} = YH^T(HH^T)^{-1} \tag{10}$$

$$\hat{\Sigma} = \frac{1}{N-p}\sum_{t=p+1}^{N}(y(t)-\hat{A}H(t))(y(t)-\hat{A}H(t))^T \tag{11}$$

โดยจะสามารถคำนวณค่า $\hat{\Sigma}$ ได้เมื่อรู้ค่า $\hat{A}$ แล้ว

6.2 การทดสอบ Wald กับปัญหากำลังสองน้อยสุด (Least-Square)

พิจารณาปัญหากำลังสองน้อยสุด (LS) เมื่อไม่มีข้อกำหนดใดๆ

$$\hat{x} = \arg\min_x \|Ax-y\|^2 = (A^TA)^{-1}A^Ty$$

เราจะสร้างข้อมูลขึ้นมาเองจากสมการ

$$y = Ax + v \tag{12}$$

โดย $y \in \mathbb{R}^N$ เมื่อ N คือจำนวนตัวอย่างของข้อมูล ให้ $x \in \mathbb{R}^n, n = 100$ หรือก็คือจำนวนพารามิเตอร์ที่ต้องการประมาณมีทั้งหมด 100 ตัว และ $\mathcal{V} \sim \mathcal{N}(0, \sigma^2 I_N)$ โดยให้ $\sigma^2 = ||x||$ และกำหนดให้ค่าจริงของ x_i บางตัวมีค่าเท่ากับศูนย์และนำไปใช้ในการทดสอบ Wald ต่อไปเพื่อพิจารณาความแม่นยำของการทดสอบ Wald ซึ่งจะพิจารณาจากค่าความผิดพลาดทั้งสองแบบของการทดสอบสมมติฐาน โดยค่าความผิดพลาดแบบที่หนึ่ง (Type I error) คือค่าความน่าจะเป็นที่ค่าพารามิเตอร์ที่*ประมาณ*ได้ไม่เท่ากับศูนย์เมื่อกำหนดให้*ค่า*จริงของพารามิเตอร์เท่ากับศูนย์

$$\text{Type I Error} = \mathbf{Prob}(\hat{x}_i \neq 0 |\, x_i = 0)$$

และค่าความผิดพลาดแบบที่สอง (Type II error)คือค่าความน่าจะเป็นที่ค่าพารามิเตอร์ที่*ประมาณ*ได้เท่ากับศูนย์เมื่อกำหนดให้*ค่า*จริงของพารามิเตอร์ไม่เท่ากับศูนย์

$$\text{Type II Error} = \mathbf{Prob}(\hat{x}_i = 0 |\, x_i \neq 0)$$

ในที่นี้จะได้ฟังก์ชันข้อกำหนดสำหรับการทดสอบ Wald ของ $\hat{x}_i$ ใดๆเมื่อ $i = 1, 2, ..., 100$ คือ

$$r(\hat{x}) = e_i^T \hat{x}$$

และสำหรับประมาณของเมทริกซ์ค่าความแปรปรวนร่วมของ $\hat{x}$ ของการประมาณกำลังสองน้อยสุดนั้นเป็นที่ทราบกันดีว่าอยู่ในรูป

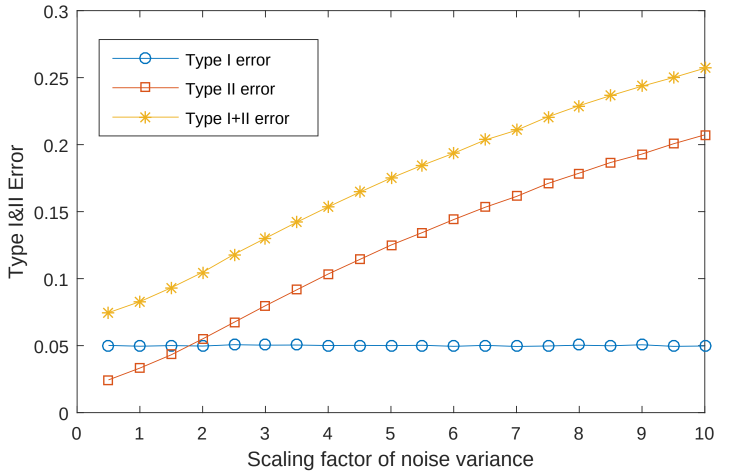
$$\widehat{\mathbf{Avar}}(\hat{x}) = \sigma^2(A^TA)^{-1} \quad \text{โดยที่ } i = 1, 2, ..., n$$

และค่า W สำหรับ $\hat{x}_i$ ใดๆจากสมการ (4) จะลดรูปเป็น

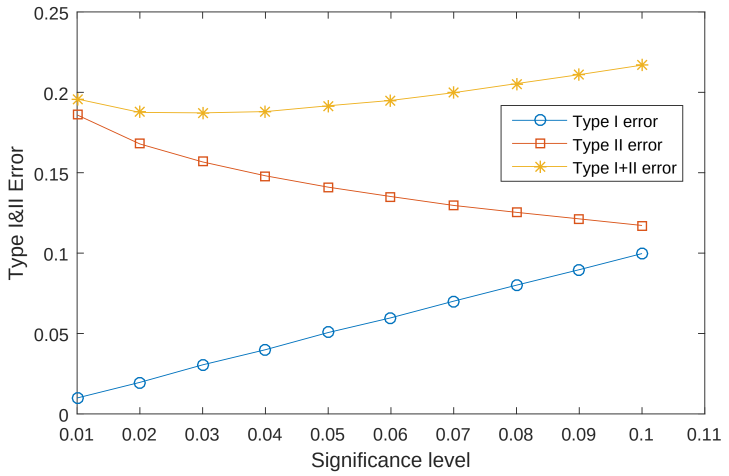
$$W_i = \hat{x}_i^2 \left[\widehat{\mathbf{Avar}}(\hat{x})_{i,i}\right]^{-1}$$

จากนั้นจึงทำการทดสอบ Wald โดยทดสอบบนสมการ (12) ที่สร้างขึ้นทั้งหมด 10,000 ครั้ง แต่ละครั้งจะคงค่า A และ x ไว้ ค่าที่เปลี่ยนไปคือสัญญาณรบกวน v เพื่อหาค่าความผิดพลาดทั้งสองแบบ และพิจารณาค่าต่างๆที่จะมีผลต่อค่าความผิดพลาดทั้งสองแบบที่เกิดขึ้น ในที่นี้จะพิจารณาค่าระดับนัยสำคัญ, ค่าความแปรปรวนของสัญญาณรบกวน และจำนวนตัวอย่างของข้อมูล

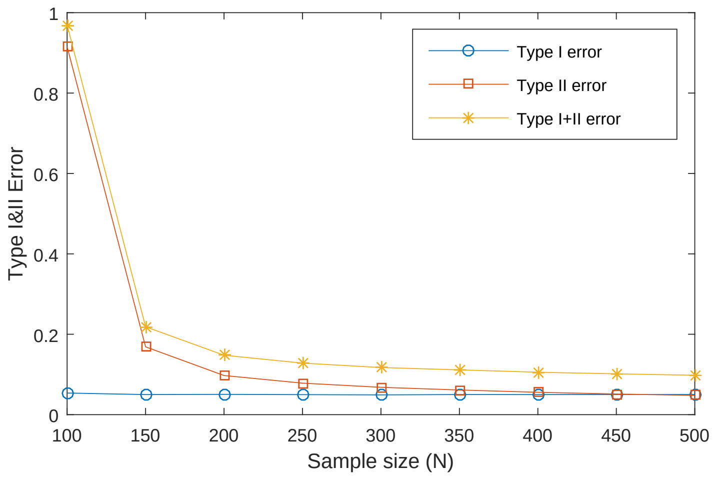
จากผลการทดลองจะเห็นว่าค่าความผิดพลาดแบบที่สองจะมีค่าลดลงเมื่อเพิ่มค่าของระดับนัยสำคัญและจำนวนตัวอย่าง แต่จะมีค่าเพิ่มมากขึ้นเมื่อค่าความแปรปรวนของสัญญาณรบกวนมีค่ามากขึ้น ในขณะที่ค่าความผิดพลาดแบบที่หนึ่งจะมีค่าประมาณเท่ากับค่าระดับนัยสำคัญเสมอ เนื่องจากการทดสอบสมมติฐานได้นิยามให้ค่าระดับนัยสำคัญและค่าความผิดพลาดแบบที่หนึ่งเป็นค่าเดียวกัน



รูป 5: รูปแสดงค่าความผิดพลาดแบบที่หนึ่ง (เส้นสีฟ้า), ค่าความผิดพลาดแบบที่สอง (เส้นสีแดง) และผลรวมค่าความผิดพลาดทั้งสองแบบ (เส้นสีเหลือง) เมื่อเปลี่ยนแปลงค่าความแปรปรวนของสัญญาณรบกวน โดย $\alpha = 0.05$, $N = 150$



รูป 6: รูปแสดงค่าความผิดพลาดแบบที่หนึ่ง (เส้นสีฟ้า), ค่าความผิดพลาดแบบที่สอง (เส้นสีแดง) และผลรวมค่าความผิดพลาดทั้งสองแบบ (เส้นสีเหลือง) เมื่อเปลี่ยนแปลงค่าระดับนัยสำคัญ (α) โดย $\sigma^2 = 10||x||$, $N = 150$



รูป 7: รูปแสดงค่าความผิดพลาดแบบที่หนึ่ง (เส้นสีฟ้า), ค่าความผิดพลาดแบบที่สอง (เส้นสีแดง) และผลรวมค่าความผิดพลาดทั้งสองแบบ (เส้นสีเหลือง) เมื่อเปลี่ยนแปลงจำนวนของตัวอย่าง โดย $\alpha = 0.05$, $\sigma^2 = 10||x||$

6.3 การทดสอบ Wald สำหรับเงื่อนไขความสัมพันธ์ระหว่างเหตุและผล

จาก (3) แบบจำลอง AR มีพารามิเตอร์ที่ต้องการประมาณค่าคือ A ซึ่งเป็นเมทริกซ์ แต่เนื่องจากการทดสอบ Wald ค่าพารามิเตอร์ที่ต้องการประมาณค่าต้องอยู่ในรูปเวคเตอร์ ดังนั้นจึงต้องเปลี่ยนรูปแบบของสมการ (3) ใหม่ โดยพารามิเตอร์ที่ต้องการประมาณของแบบจำลองคือ

$$\theta = \begin{bmatrix} B_{11}^T & \cdots & B_{1n}^T & B_{21}^T & \cdots & B_{2n}^T & \cdots & B_{nn}^T \end{bmatrix}^T \in \mathbb{R}^{n^2 p} \tag{13}$$

เมื่อ $B_{ij} = ((A_1)_{ij}, (A_2)_{ij}, ..., (A_p)_{ij}) \in \mathbb{R}^p$ ดังนั้นแบบจำลอง AR จาก (3) จะเปลี่ยนรูปเป็นรูปแบบของสมการใหม่คือ

$$y(t) = \bar{H}(t)\theta + v(t) \tag{14}$$

เมื่อกำหนดให้ $\bar{y}_i = \begin{bmatrix} y_i(t-1) \\ y_i(t-2) \\ \vdots \\ y_i(t-p) \end{bmatrix}_{p\times 1}$ และ $\bar{y} = \begin{bmatrix} \bar{y}_1 \\ \bar{y}_2 \\ \vdots \\ \bar{y}_n \end{bmatrix}_{np\times 1}$ จะได้ $\bar{H}(t) = \begin{bmatrix} \bar{y}^T & 0 & 0 & \cdots & 0 \\ 0 & \bar{y}^T & 0 & \cdots & 0 \\ 0 & 0 & \bar{y}^T & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \bar{y}^T \end{bmatrix}_{n\times n^2 p}$

จากนั้นจึงใช้การทดสอบ Wald สำหรับทดสอบ $(A_k)_{ij}$ เมื่อ $k = 1, 2, ..., p$ โดยทดสอบทีละคู่ (i, j) สำหรับ $j > i, i = 1, 2, ..., n$ ทั้งหมด $n^2 - n$ ครั้ง โดยฟังก์ชันข้อกำหนดในการทดสอบ Wald สำหรับแต่ละคู่ (i, j) คือ

$$r(\theta) = \begin{bmatrix} (A_1)_{ij} \\ (A_2)_{ij} \\ \vdots \\ (A_p)_{ij} \end{bmatrix} \tag{15}$$

หลังจากทำการทดสอบ Wald จะสามารถหาโครงสร้างเชิงสาเหตุของพารามิเตอร์ A ของแบบจำลอง AR ได้ออกมาในรูปแบบโครงสร้างของศูนย์ซึ่งจะถูกกล่าวถึงในหัวข้อที่ 4.2