การเปรียบเทียบสถิติทดสอบความเท่ากันของเมทริกซ์ความแปรปรวนร่วมสองประชากรสำหรับข้อมูลที่มีมิติสูง

A Comparison of Tests for Equality of Two Population Covariance Matrices for High-Dimensional Data

ศศิภรณ์ สิทธิศร*, เสาวภา ชัยพิทักษ์ และธิดาพร ศุภภากร

ภาควิชาสถิติ คณะวิทยาศาสตร์ มหาวิทยาลัยเกษตรศาสตร์ วิทยาเขตบางเขน

แขวงลาดยาว เขตจตุจักร กรุงเทพมหานคร 10900

Sasiporn Sitthisorn*, Saowapa Chaipitak and Thidaporn Supapakorn

Department of Statistics, Faculty of Science, Kasetsart University, Bangkhen Campus,

Ladyao, Chatuchak, Bangkok 10900

บทคัดย่อ

งานวิจัยนี้มีวัตถุประสงค์เพื่อ เปรียบเทียบประสิทธิภาพของสถิติทดสอบความเท่ากันของเมทริกซ์ความแปรปรวนร่วมสองประชากร สำหรับข้อมูลที่มีการแจกแจงปรกติหลายตัวแปรและเป็นข้อมูลที่มิติสูง (high-dimensional data) ของสถิติทดสอบ 2 สถิติ ได้แก่ สถิติทดสอบ LC ของ Li และ Chen และสถิติทดสอบ SE ของ Srivastava และคณะ โดยมีเกณฑ์ในการเปรียบเทียบ คือ ค่าความน่าจะเป็นของความคลาดเคลื่อนประเภทที่ 1 และอำนาจการทดสอบ จำลองข้อมูลภายใต้สถานการณ์ต่าง ๆ ด้วยเทคนิคมอนติคาร์โล กระทำซ้ำ 1,000 ครั้ง โดยกำหนดให้ข้อมูลมีการแจกแจงปรกติหลายตัวแปรและมีโครงสร้างของเมทริกซ์ ความแปรปรวนร่วมของประชากร 5 แบบ ได้แก่ โครงสร้าง compound symmetry, simple, Toeplitz, unstructured และโครงสร้าง variance components และกำหนดให้ขนาดตัวอย่างทั้งสองกลุ่มเท่ากัน ($n_1=n_2$) และจำนวนตัวแปรมีมากกว่าหรือเท่ากับขนาดตัวอย่าง ดังนี้ $(n_1=n_2,p) = (10,10),(10,20),(10,30),(20,20),(20,40),(20,60),(40,40),(40,80),(40,120)$ ผลการวิจัยพบว่าสถิติทดสอบ LC สามารถควบคุมความน่าจะเป็นของความคลาดเคลื่อนประเภทที่ 1 ได้ทุกโครงสร้างของเมทริกซ์ความแปรปรวนร่วมที่พิจารณา ในขณะที่สถิติทดสอบ SE ไม่สามารถควบคุมความน่าจะเป็นของความคลาดเคลื่อนประเภทที่ 1 ได้ นอกจากนี้สถิติทดสอบ LC มีค่าอำนาจการทดสอบสูงกว่าสถิติทดสอบ SE และลู่เข้าสู่ 1 เมื่อเพิ่มขนาดตัวอย่างและจำนวนตัวแปรในทุกสถานการณ์

คำสำคัญ : การแจกแจงปรกติหลายตัวแปร; ความน่าจะเป็นของความคลาดเคลื่อนประเภทที่ 1; อำนาจการทดสอบ; เมทริกซ์ความแปรปรวนร่วม

***ผู้รับผิดชอบบทความ :** sitthisorn.ssp@gmail.com

Abstract

The objective of this research is to compare the efficiency between two test statistics: Li and Chen's test statistic and Srivastava *et al.*'s test statistic, for testing the equality of two covariance matrices for high-dimensional data distributed as multivariate normal (p variables). There are two criteria for comparing the tests consisting of the probability of type I error and power of the test which were measured through data simulation using Monte Carlo technique 1000 iterations. The two high-dimensional data distributed as multivariate normal data (p variables) under five covariance matrix structures: compound symmetry, simple, Toeplitz, unstructured, and variance components, were simulated. The number of variables (p) was assigned to be greater than or equal to its sample sizes and varied in $(n_1=n_2, p) = (10,10),(10,20),(10,30),(20,20),(20,40),(20,60),$ $(40,40),(40,80),(40,120)$. It was shown that Li and Chen's test statistic can control the probability of type I error for all covariance matrix structures considered whereas Srivastava et al. is unable. In addition, the Li and Chen's test statistic was strongly higher power than the other one and its power converged to one when p and sample sizes were increased for all situations.

Keywords: multivariate normal distribution; probability of type I error; power of the test; covariance matrix

1. บทนำ

ในการวิเคราะห์ข้อมูลหลายตัวแปร (multivariate analysis) เมื่อต้องการทดสอบความเท่ากันของเวกเตอร์ค่าเฉลี่ยสองประชากร สำหรับข้อมูลมาจากประชากรที่มีการแจกแจงปรกติหลายตัวแปร ภายใต้ความเท่ากันของเมทริกซ์ความแปรปรวนร่วมของสองประชากร สถิติทดสอบที่เหมาะสมและนิยมใช้ คือ Hotelling's T^2 แต่ในกรณีที่เมทริกซ์ความแปรปรวนร่วมของสองประชากรไม่เท่ากัน สถิติทดสอบที่นิยมใช้ คือ Behrens-Fisher [1] ในการทดสอบความเท่ากันของเมทริกซ์ความแปรปรวนร่วมของสองประชากรโดยปกติแล้วอาศัยการทดสอบของบ็อกซ์ (Box's test) [2] และการทดสอบอัตราส่วนภาวะน่าจะเป็น (Likelihood ratio test, LRT) [3] เมื่อข้อมูลที่มีจำนวนตัวแปร (p) น้อยกว่าขนาดตัวอย่างทั้งสองกลุ่ม (n_1 และ n_2) สำหรับกรณีที่ข้อมูลที่มีจำนวนตัวแปรมากกว่าหรือเท่ากับขนาดตัวอย่าง กรณีเช่นนี้เรียกว่าข้อมูลที่มีมิติสูง (high-dimensional data) ข้อมูลที่มีลักษณะนี้จะไม่สามารถคำนวณเมทริกซ์ผกผันของเมทริกซ์ความแปรปรวนร่วมของตัวอย่างได้ ทำให้การทดสอบของบ็อกซ์และ LRT ประสบปัญหาไม่สามารถคำนวณค่าได้ [2] จึงมีผู้ศึกษาและเสนอวิธีแก้ปัญหาเกี่ยวกับการทดสอบความเท่ากันของเมทริกซ์ความแปรปรวนร่วมของสองประชากรสำหรับข้อมูลที่มีมิติสูง โดยในปี ค.ศ. 2007 Schott [4] ได้นำเสนอสถิติทดสอบความเท่ากันของเมทริกซ์ความแปรปรวนร่วมของหลายประชากรที่มีการแจกแจงปรกติหลายตัวแปรสำหรับข้อมูลที่มีมิติสูง ซึ่งสถิติทดสอบที่นำเสนอได้พัฒนาขึ้นมาจากแนวคิดจากหาค่าของผลรวมของกำลังสองของผลต่างของเมทริกซ์ความแปรปรวนร่วมของสองประชากร สถิติทดสอบของ Schott มีการแจกแจงเข้าสู่การแจกแจงปรกติมาตรฐาน เมื่อ $p > n_1, n_2$

และ $(n_1, n_2, p) \to \infty$ เมื่อเปรียบเทียบประสิทธิภาพของสถิติทดสอบของ Schott กับสถิติทดสอบ LRT พบว่าสถิติทดสอบของ Schott มีค่าอำนาจการทดสอบสูงกว่าสถิติทดสอบ LRT และในปีเดียวกัน Srivastava [5] ได้คิดค้นสถิติทดสอบความเท่ากันสำหรับเมทริกซ์ความแปรปรวนร่วมของหลายประชากรที่มีการแจกแจงปรกติหลายตัวแปรสำหรับข้อมูลที่มีมิติสูง โดยพัฒนาขึ้นมาจากแนวคิดการประมาณค่าขอบล่างนอร์มของกำลังสองของผลต่างของเมทริกซ์ของความแปรปรวนร่วมของสองประชากร และสถิติทดสอบของ Srivastava มีการแจกแจงเข้าสู่การแจกแจงปรกติมาตรฐานเมื่อ $p > n_1, n_2$ และ $(n_1, n_2, p) \to \infty$ ในปี ค.ศ. 2010 Srivastava และ Yanagahara [6] ได้นำเสนอสถิติทดสอบความเท่ากันของเมทริกซ์ความแปรปรวนร่วมของสองประชากรที่มีการแจกแจงปรกติหลายตัวแปรสำหรับข้อมูลที่มีมิติสูง โดยพัฒนาจากผลต่างของผลรวมของเมทริกซ์ความแปรปรวนร่วมกำลังสองของสองประชากร และเปรียบเทียบประสิทธิภาพของสถิติทดสอบที่นำเสนอกับสถิติทดสอบ 2 ตัวทดสอบ คือ สถิติทดสอบของ Schott [4] และสถิติทดสอบของ Srivastava [5] เกณฑ์ที่ใช้ในการเปรียบเทียบประสิทธิภาพ คือ ความน่าจะเป็นของความคลาดเคลื่อนประเภทที่ 1 และอำนาจการทดสอบที่ได้จากการจำลองข้อมูล ผลการวิจัยสรุปได้ดังนี้ สถิติทดสอบที่นำเสนอมีค่าอำนาจการทดสอบสูงกว่าสถิติทดสอบของ Schott [4] และสถิติทดสอบของ Srivastava [5] และค่าความน่าจะเป็นของความคลาดเคลื่อนประเภทที่ 1 สถิติทดสอบของ Srivastava [5] มีประสิทธิภาพดีที่สุดทุกขนาดตัวอย่างที่กำหนด ต่อมาในปี ค.ศ. 2012 Li และ Chen [7] นำเสนอสถิติทดสอบความเท่ากันของเมทริกซ์ความแปรปรวนร่วมสองประชากรที่มีการแจกแจงแบบใด ๆ สำหรับข้อมูลที่มีมิติสูง และเปรียบเทียบประสิทธิภาพของสถิติทดสอบที่นำเสนอกับสถิติทดสอบ LRT และสถิติทดสอบอัตราส่วนภาวะน่าจะเป็นปรับแก้ (Corrected likelihood ratio test, CLRT) ซึ่งนำเสนอโดย Bai และคณะ ในปี ค.ศ. 2009 [8] ผลการศึกษาพบว่าสถิติทดสอบที่นำเสนอและสถิติทดสอบ CLRT มีค่าอำนาจการทดสอบสูงกว่าสถิติทดสอบ LRT และมีความน่าจะเป็นของความคลาดเคลื่อนประเภทที่ 1 ประสิทธิภาพดีที่สุดทุกขนาดตัวอย่างจากข้อมูลที่มีมิติสูง และในปี ค.ศ. 2014 Srivastava และคณะ [9] ได้นำเสนอสถิติทดสอบสำหรับทดสอบเมทริกซ์ความแปรปรวนร่วมของหนึ่ง และสองประชากรที่มีการแจกแจงปรกติหลายตัวแปรสำหรับข้อมูลที่มีมิติสูง เกณฑ์ที่ใช้ในการเปรียบเทียบประสิทธิภาพคือ ความน่าจะเป็นของความคลาดเคลื่อนประเภทที่ 1 และอำนาจการทดสอบสำหรับหนึ่งประชากรได้เปรียบเทียบประสิทธิภาพของสถิติทดสอบกับสถิติทดสอบของ Srivastava [5] พบว่าสถิติทดสอบที่นำเสนอมีประสิทธิภาพกว่า และสำหรับสองประชากรได้เปรียบเทียบประสิทธิภาพของสถิติทดสอบกับสถิติทดสอบของ Schott [4] พบว่าสถิติทดสอบที่นำเสนอประสิทธิภาพกว่า

ดังนั้นในครั้งนี้ผู้วิจัยสนใจนำวิธีการเปรียบเทียบสถิติทดสอบความเท่ากันของเมทริกซ์ความแปรปรวนร่วมสองประชากรสำหรับข้อมูลที่มีมิติสูง 2 วิธี ได้แก่ สถิติทดสอบ LC ของ Li และ Chen และสถิติทดสอบ SE ของ Srivastava และคณะ โดยกำหนดโครงสร้างของเมทริกซ์ความแปรปรวนของประชากรเป็น 5 แบบ ได้แก่ โครงสร้าง compound symmetry โครงสร้าง simple โครงสร้าง Toeplitz โครงสร้าง unstructured และ โครงสร้าง variance components เพื่อให้ได้สารสนเทศประกอบการตัดสินใจในการเลือกใช้สถิติทดสอบที่เหมาะสมในแต่ละสถานการณ์ และเหมาะสมกับการวิเคราะห์ข้อมูลที่มีมิติสูงต่อไป

2. วิธีการวิจัย

2.1 จำลองข้อมูล $\boldsymbol{x}_{ij}$ ด้วยโปรแกรม R โดยกำหนดให้ $\boldsymbol{x}_{ij} = \left(x_{ij1}, x_{ij2}, ..., x_{ijp}\right)'$ เมื่อ $i = 1, 2$ และ $j = 1,2,...,n_i$ เป็นตัวอย่างสุ่มจากประชากรที่ $\boldsymbol{i}$ ที่มีการแจกแจงปรกติหลายตัวแปร ($\boldsymbol{p}$ ตัวแปร) โดยมีเวกเตอร์ค่าเฉลี่ย $\boldsymbol{\mu}_i$ และเมทริกซ์ความแปรปรวนร่วม Σ_i หรือ $\boldsymbol{x}_{ij} \sim N_p\left(\boldsymbol{\mu}_i, \Sigma_i\right)$ โดยกำหนดสถานการณ์ที่มีขนาดตัวอย่าง n_1 และ n_2 และจำนวนตัวแปร p ที่ใช้ในการศึกษาคือ (n_1=n_2, p) = (10,10), (10,20), (20,20), (20,40), (20,60), (40,40), (40,80), (40,120) ตามลำดับ และมีโครงสร้างของเมทริกซ์ความแปรปรวนร่วม [10] ภายใต้สมมติฐานดังต่อไปนี้

2.1.1 โครงสร้าง compound symmetry (CS) ภายใต้สมมติฐาน คือ

$$H_0^1 : \Sigma_1 = \Sigma_2 = \boldsymbol{C}_0$$

$$H_1^1 : \Sigma_1 = \boldsymbol{C}_0 \text{ และ } \Sigma_2 = \boldsymbol{C}_1$$

เมื่อ $\boldsymbol{C}_0 = 0.99\boldsymbol{I}_p + 0.01\boldsymbol{I}_p\boldsymbol{I}_p'$ และ $\boldsymbol{C}_1 = 0.90\boldsymbol{I}_p + 0.10\boldsymbol{I}_p\boldsymbol{I}_p'$

2.1.2 โครงสร้าง simple (SIM) ภายใต้สมมติฐาน คือ

$$H_0^2 : \Sigma_1 = \Sigma_2 = 2\boldsymbol{I}_p$$

$$H_1^2 : \Sigma_1 = 2\boldsymbol{I}_p \text{ และ } \Sigma_2 = 4\boldsymbol{I}_p$$

2.1.3 โครงสร้าง Toeplitz (TOEP) ภายใต้สมมติฐาน คือ

$$H_0^3 : \Sigma_1 = \Sigma_2 = \boldsymbol{T}_0$$

$$H_1^3 : \Sigma_1 = \boldsymbol{T}_0 \text{ และ } \Sigma_2 = \boldsymbol{T}_1$$

เมื่อ $\boldsymbol{T}_0 = \left(\sigma_{ij}\right)_{p\times p} = \begin{cases} 1 & ;|i-j| = 0 \\ -0.5 & ;|i-j| = 1 \\ 0 & ;|i-j| \neq 0,1 \end{cases}$

และ $\boldsymbol{T}_1 = \left(\sigma_{ij}\right)_{p\times p} = \begin{cases} 1 & ;|i-j| = 0 \\ -0.05 & ;|i-j| = 1 \\ 0 & ;|i-j| \neq 0,1 \end{cases}$

2.1.4 โครงสร้าง unstructured (UN) ภายใต้สมมติฐาน คือ

$$H_0^4 : \Sigma_1 = \Sigma_2 = \boldsymbol{U}_0$$

$$H_1^4 : \Sigma_1 = \boldsymbol{U}_0 \text{ และ } \Sigma_2 = \boldsymbol{U}_1$$

เมื่อ $\boldsymbol{U}_0 = \left(\sigma_{ij}\right)_{p\times p} = \begin{cases} 1 & ; i = j \\ \dfrac{(-1)^{i+j}(0.1)i}{j} & ; i \neq j \end{cases}$

และ $\boldsymbol{U}_1 = \left(\sigma_{ij}\right)_{p\times p} = \begin{cases} 1 & ; i = j \\ \dfrac{(-1)^{i+j}(0.5)i}{j} & ; i \neq j \end{cases}$

2.1.5 โครงสร้าง variance components (VC) ภายใต้สมมติฐานคือ

$$H_0^5 : \Sigma_1 = \Sigma_2 = \boldsymbol{A}_0$$

$$H_1^5 : \Sigma_1 = \boldsymbol{A}_0 \text{ และ } \Sigma_2 = \boldsymbol{A}_1$$

เมื่อ $\boldsymbol{A}_0 = diag\left(\sigma_{11}, \sigma_{22}, ..., \sigma_{pp}\right)$ และ $\sigma_{ij} \sim U(1,2), i = 1,...,p$

และ $\boldsymbol{A}_1 = diag\left(\sigma_{11}, \sigma_{22}, ..., \sigma_{pp}\right)$ และ $\sigma_{ij} \sim U(3,4), i = 1,...,p$

2.2 คำนวณค่าสถิติทดสอบความเท่ากันของเมทริกซ์ความแปรปรวนร่วมสองประชากรสำหรับข้อมูลที่มีมิติสูงสำหรับ 2 สถิติทดสอบ โดยมีวิธีการคำนวณดังต่อไปนี้

2.2.1 สถิติทดสอบ LC ของ Li และ Chen โดย Li และ Chen นำเสนอสถิติทดสอบความเท่ากันของเมทริกซ์ความแปรปรวนร่วมสองประชากร โดยหาตัวประมาณของ $tr\left(\Sigma_1 - \Sigma_2\right)^2 = tr\left(\Sigma_1^2\right) + tr\left(\Sigma_2^2\right) - 2tr\left(\Sigma_1\Sigma_2\right)$ และได้สถิติทดสอบ LC ดังนี้

$$LC = \frac{A_{n_1} + A_{n_2} - 2C_{n_1,n_2}}{\dfrac{2A_{n_1}}{n_2} + \dfrac{2A_{n_2}}{n_1}} \qquad (1)$$

โดยที่ $A_{n_i} = \dfrac{1}{n_i(n_i-1)}\sum_{j\neq k}^{n_i}\left(\boldsymbol{X}_{ij}'\boldsymbol{X}_{ik}\right)^2 - \dfrac{2}{n_i(n_i-1)(n_i-2)}\sum_{j\neq k\neq m}^{n_i}\boldsymbol{X}_{ij}'\boldsymbol{X}_{ik}\boldsymbol{X}_{ik}'\boldsymbol{X}_{im} + \dfrac{1}{n_i(n_i-1)(n_i-2)(n_i-3)}\sum_{j\neq k\neq m\neq n}^{n_i}\boldsymbol{X}_{ij}'\boldsymbol{X}_{ik}\boldsymbol{X}_{im}'\boldsymbol{X}_{in}$

$$C_{n_1,n_2} = \frac{1}{n_1n_2}\sum_{j=1}^{n_1}\sum_{k=1}^{n_2}\left(\boldsymbol{X}_{1j}'\boldsymbol{X}_{2k}\right)^2 - \frac{1}{n_1n_2(n_1-1)}\sum_{j\neq m}^{n_1}\sum_{k}^{n_2}\boldsymbol{X}_{1j}'\boldsymbol{X}_{2k}\boldsymbol{X}_{2k}'\boldsymbol{X}_{1m} - \frac{1}{n_1n_2(n_2-1)}\sum_{j\neq m}^{n_2}\sum_{k=1}^{n_1}\boldsymbol{X}_{2j}'\boldsymbol{X}_{1k}\boldsymbol{X}_{1k}'\boldsymbol{X}_{2m} + \frac{1}{n_1n_2(n_1-1)(n_2-1)}\sum_{j\neq m}^{n_2}\sum_{k\neq n}^{n_1}\boldsymbol{X}_{1j}'\boldsymbol{X}_{2k}\boldsymbol{X}_{1k}'\boldsymbol{X}_{2m}$$

เมื่อ A_{n_i} เป็นตัวประมาณไม่เอนเอียงของ $tr\left(\Sigma_i^2\right)$ และ C_{n_1,n_2} เป็นตัวประมาณไม่เอนเอียงของ $tr\left(\Sigma_1 \Sigma_2\right)$

สถิติทดสอบ LC มีการแจกแจงเข้าสู่การแจกแจงปรกติมาตรฐาน เมื่อ $(n_1, n_2, p) \to \infty$ และปฏิเสธสมมติฐานหลัก เมื่อค่าสถิติทดสอบ LC ที่คำนวณได้มีค่ามากกว่าค่าวิกฤตจากตารางการแจกแจงปรกติมาตรฐานที่ระดับนัยสำคัญ α หรือ LC $> Z_{1-\alpha}$ เมื่อ $Z_{1-\alpha}$ คือ ค่าคะแนนมาตรฐาน ณ เปอร์เซ็นต์ไทล์ที่ $(1-\alpha)100\%$ ของตัวแปรสุ่ม Z ที่มีการแจกแจงปรกติมาตรฐาน

2.2.2 สถิติทดสอบ SE ของ Srivastava และคณะ โดย Srivastava และคณะ พัฒนาสถิติทดสอบของ Schott และได้นำเสนอสถิติทดสอบความเท่ากันของเมทริกซ์ความแปรปรวนร่วมสองประชากร ค่าสถิติทดสอบ SE คำนวณดังนี้

$$SE = \frac{\hat{a}_{21} - \hat{a}_{22} - 2tr\left(\boldsymbol{V}_1\boldsymbol{V}_2\right) / p\left(n_1 - 1\right)\left(n_2 - 1\right)}{2\hat{a}_2\left(1/\left(n_1-1\right) + 1/\left(n_2-1\right)\right)} \qquad (2)$$

โดยที่ $\hat{a}_{2i} = \frac{1}{f_i}\left\{\left(n_i - 2\right)n_i tr\left[\boldsymbol{V}_i^2\right] - n(n-1)tr\left[\boldsymbol{D}_i^2\right] + tr\left[\boldsymbol{V}_i^2\right]\right\}$

เมื่อ $f_i = pn_i(n_i-1)(n_i-2)(n_i-3)$ $\boldsymbol{D}_i = diag\left(\boldsymbol{Y}_{ij}'\boldsymbol{Y}_{ij}, ..., \boldsymbol{Y}_{in_i}'\boldsymbol{Y}_{in_i}\right)$

; $\boldsymbol{Y}_{ij} = \boldsymbol{X}_{ij} - \bar{\boldsymbol{X}}_i$ และ $\boldsymbol{V}_i = \sum_{j=1}^{n_i}\left(\boldsymbol{X}_{ij} - \bar{\boldsymbol{X}}_i\right)\left(\boldsymbol{X}_{ij} - \bar{\boldsymbol{X}}_i\right)'$

สถิติทดสอบ SE มีการแจกแจงเข้าสู่การแจกแจงปรกติมาตรฐาน เมื่อ $(n_1, n_2, p) \to \infty$ และปฏิเสธสมมติฐานหลัก เมื่อค่าสถิติทดสอบ SE ที่คำนวณได้มีค่ามากกว่าค่าวิกฤต $Z_{1-\alpha}$ จากตารางการแจกแจงปรกติมาตรฐาน ที่ระดับนัยสำคัญ α หรือ SE $> Z_{1-\alpha}$

2.3 นำค่าสถิติทดสอบที่คำนวณได้มาเปรียบเทียบกับค่าวิกฤตในแต่ละตัวสถิติทดสอบ และทดลองซ้ำจำนวน 1,000 รอบ ในแต่ละสถานการณ์เพื่อคำนวณหาค่าความน่าจะเป็นของความคลาดเคลื่อนประเภทที่ 1 โดยนับจำนวนครั้งของการปฏิเสธสมมติฐานหลัก (H_0) เมื่อสมมติฐานหลักเป็นจริง แล้วหารด้วยจำนวนรอบทั้งหมด 1,000 รอบ จากเกณฑ์ของ Cochran ปี ค.ศ. 1954 [11] ที่ระดับนัยสำคัญ 0.05 ค่าความน่าจะเป็นของความคลาดเคลื่อนประเภทที่ 1 ที่คำนวณได้ควรอยู่ในช่วง [0.040, 0.060]

2.4 คำนวณหาค่าอำนาจการทดสอบ โดยกำหนดให้ข้อมูลมีโครงสร้างของเมทริกซ์ความแปรปรวนร่วมตามสมมติฐานรอง (H_1) ของแต่ละโครงสร้างของเมทริกซ์ความแปรปรวนร่วม และคำนวณค่าสถิติทดสอบทั้งสอง กระทำซ้ำจำนวน 1,000 รอบ และนับจำนวนครั้งของการปฏิเสธสมมติฐานหลัก (H_0) แล้วหารด้วยจำนวนรอบทั้งหมด 1,000 รอบ

3. ผลการวิจัย

การวิจัยครั้งนี้มีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพของสถิติทดสอบความเท่ากันของเมทริกซ์ความแปรปรวนร่วมของสองประชากรสำหรับข้อมูลที่มีมิติสูง โดยศึกษาสถิติทดสอบ 2 วิธี คือ สถิติทดสอบ LC และสถิติทดสอบ SE โดยพิจารณาประสิทธิภาพของสถิติทดสอบจากความสามารถในการควบคุมความน่าจะเป็นของความคลาดเคลื่อนประเภทที่ 1 และอำนาจการทดสอบ เมื่อกำหนดให้ขนาดของตัวอย่างเท่ากันและมีโครงสร้างของเมทริกซ์ความแปรปรวนร่วมของประชากรเป็น 5 แบบ โดยจำลองข้อมูลด้วยวิธีมอนติคาร์โลในแต่ละสถานการณ์ทำซ้ำ 1,000 ครั้ง ได้ค่าความน่าจะเป็นของความคลาดเคลื่อนประเภทที่ 1 และอำนาจการทดสอบแสดงดังตารางที่ 1 และ 2

ตารางที่ 1 พบว่าเมื่อกำหนดให้ประชากรทั้งสองกลุ่มมีความแปรปรวนร่วมเท่ากัน $(\Sigma_1 = \Sigma_2)$ ตาม H_0^1 โดยมีโครงสร้าง compound symmetry พบว่าค่าความน่าจะเป็นของความคลาดเคลื่อนประเภทที่ 1 ของสถิติทดสอบ LC มีค่าอยู่ในช่วง [0.040,0.060] ตามเกณฑ์ของ Cochran และเข้าใกล้ระดับนัยสำคัญ

0.05 นั่นแสดงว่าสถิติทดสอบ LC สามารถควบคุมความน่าจะเป็นของความคลาดเคลื่อนประเภทที่ 1 ได้ทุกกรณีของ n_1, n_2 และ p ในขณะที่สถิติทดสอบ SE มีค่าความน่าจะเป็นของความคลาดเคลื่อนประเภทที่ 1 เข้าใกล้ศูนย์ นั่นหมายความว่า SE ไม่สามารถควบคุมความน่าจะเป็นของความคลาดเคลื่อนประเภทที่ 1 ได้

ตารางที่ 1 ค่าความน่าจะเป็นของความคลาดเคลื่อนประเภทที่ 1 ของสถิติทดสอบ LC และ SE จำแนกตามโครงสร้างของเมทริกซ์ความแปรปรวนร่วม ที่ระดับนัยสำคัญ $\alpha = 0.05$

ขนาดตัวอย่าง ($n_1 = n_2$)	จำนวนตัวแปร (p)	โครงสร้างของเมทริกซ์ความแปรปรวนร่วม									
		CS		SIM		TOEP		UN		VC	
		LC	SE	LC	SE	LC	SE	LC	SE	LC	SE
10	10	0.058	0.014	0.044	0.014	0.049	0.015	0.054	0.014	0.052	0.010
	20	0.059	0.008	0.048	0.007	0.041	0.006	0.049	0.007	0.045	0.004
	30	0.056	0.001	0.054	0.001	0.050	0.001	0.047	0.002	0.040	0.001
20	20	0.055	0.019	0.058	0.018	0.057	0.018	0.053	0.016	0.049	0.017
	40	0.052	0.006	0.067	0.006	0.056	0.006	0.051	0.007	0.059	0.007
	60	0.058	0.001	0.063	0.001	0.051	0.001	0.054	0.002	0.049	0.001
40	40	0.060	0.012	0.051	0.013	0.057	0.013	0.043	0.019	0.056	0.012
	80	0.053	0.006	0.053	0.007	0.046	0.008	0.047	0.010	0.040	0.008
	120	0.054	0.001	0.045	0.001	0.051	0.001	0.052	0.008	0.043	0.001

ในกรณีที่กำหนดให้ประชากรทั้งสองกลุ่มมีความแปรปรวนร่วมเท่ากันตาม H_0^2 ตามโครงสร้าง simple พบว่าสถิติทดสอบ LC สามารถควบคุมความน่าจะเป็นของความคลาดเคลื่อนประเภทที่ 1 ได้ทุกกรณีของ n_1, n_2 และ p เนื่องจากค่าความน่าจะเป็นของความคลาดเคลื่อนประเภทที่ 1 มีค่าเข้าใกล้ระดับนัยสำคัญ 0.05 ในขณะที่สถิติทดสอบ SE ไม่สามารถควบคุมความน่าจะเป็นของความคลาดเคลื่อนประเภทที่ 1 ได้

สำหรับกรณีที่กำหนดให้ประชากรทั้งสองกลุ่มมีความแปรปรวนร่วมเท่ากันโดยตาม H_0^3 ตามโครงสร้าง Toeplitz พบว่าสถิติทดสอบ LC มีค่าความน่าจะเป็นของความลาดเคลื่อนประเภทที่ 1 ส่วนใหญ่มีค่าเข้าใกล้ระดับนัยสำคัญ 0.05 ซึ่งสามารถควบคุมความน่าจะเป็นของความคลาดเคลื่อนประเภทที่ 1 ได้ดีกว่าสถิติทดสอบ SE เนื่องจากสถิติทดสอบ SE ส่วนใหญ่มีค่าความน่าจะเป็นของความคลาดเคลื่อนประเภทที่ 1 แตกต่างจากระดับนัยสำคัญ 0.05 และมีค่าใกล้ศูนย์

ในกรณีที่กำหนดให้ประชากรทั้งสองกลุ่มมีความแปรปรวนร่วมเท่ากันตาม H_0^4 ตามโครงสร้าง unstructured พบว่าสถิติทดสอบ LC สามารถควบคุมความน่าจะเป็นของความคลาดเคลื่อนประเภทที่ 1 เนื่องจากค่าความน่าจะเป็นของความคลาดเคลื่อนประเภทที่ 1 เข้าใกล้ระดับนัยสำคัญ 0.05 ทุกกรณีของค่า n_1, n_2 และ p ในขณะที่สถิติทดสอบ SE ไม่สามารถควบคุมความน่าจะเป็นของความคลาดเคลื่อน

ประเภทที่ 1 ได้ เนื่องจากมีค่าความน่าจะเป็นของความคลาดเคลื่อนประเภทที่ 1 มีค่าเข้าใกล้ศูนย์

ในทำนองเดียวกัน เมื่อกำหนดให้ประชากรทั้งสองกลุ่มมีความแปรปรวนร่วมเท่ากันตาม H_0^5 ตามโครงสร้าง variance component พบว่าค่าความน่าจะเป็นของความคลาดเคลื่อนประเภทที่ 1 ของสถิติทดสอบ LC ส่วนใหญ่มีค่าเข้าใกล้ระดับนัยสำคัญ 0.05 เมื่อเพิ่มขนาดตัวอย่างและเพิ่มจำนวนตัวแปร นั่นหมายความว่าสถิติทดสอบ LC สามารถควบคุมความน่าจะเป็นของความคลาดเคลื่อนประเภท ที่ 1 ในขณะที่สถิติทดสอบ SE มีค่าความน่าจะเป็นของความคลาดเคลื่อนประเภทที่ 1 เข้าใกล้ศูนย์ สถิติทดสอบ SE จึงไม่สามารถควบคุมความน่าจะเป็นของความคลาดเคลื่อนประเภทที่ 1 ได้ทุกกรณีของ n_1, n_2 และ p

เมื่อพิจารณาอำนาจการทดสอบจากตารางที่ 2 พบว่าเมื่อกำหนดให้ประชากรทั้งสองกลุ่มมีความแปรปรวนร่วมแตกต่างกัน $(\Sigma_1 \neq \Sigma_2)$ ตาม H_1^1 (โครงสร้าง compound symmetry) หรือ H_1^2 (โครงสร้าง Simple) หรือ H_1^3 (โครงสร้าง Toeplitz) หรือ H_1^4 (โครงสร้าง unstructured) หรือ H_1^5 (โครงสร้าง variance components) พบว่าทุกโครงสร้างให้ผลทำนองเดียวกัน คือ สถิติทดสอบ LC ให้ค่าอำนาจการทดสอบสูงกว่าสถิติทดสอบ SE และมีค่าลู่เข้าสู่ 1 เมื่อ n_1, n_2 และ p มีค่ามากขึ้น

ตารางที่ 2 ค่าอำนาจการทดสอบของสถิติทดสอบ LC และ SE จำแนกตามโครงสร้างของเมทริกซ์ความแปรปรวนร่วม

ขนาดตัวอย่าง	จำนวนตัวแปร	โครงสร้างของเมทริกซ์ความแปรปรวนร่วม									
		CS		SIM		TOEP		UN		VC	
($n_1 = n_2$)	(p)	LC	SE	LC	SE	LC	SE	LC	SE	LC	SE
10	10	0.060	0.028	0.234	0.104	0.071	0.092	0.281	0.111	0.107	0.133
	20	0.069	0.011	0.231	0.037	0.065	0.037	0.320	0.176	0.195	0.109
	30	0.043	0.013	0.236	0.010	0.057	0.014	0.404	0.227	0.153	0.021
20	20	0.166	0.088	0.638	0.427	0.523	0.299	0.630	0.525	0.360	0.706
	40	0.340	0.159	0.631	0.247	0.445	0.173	0.814	0.723	0.870	0.550
	60	0.515	0.220	0.634	0.126	0.578	0.081	0.891	0.825	0.890	0.377
40	40	0.852	0.584	0.916	0.950	0.931	0.845	0.909	0.989	0.875	0.990
	80	0.777	0.755	0.956	0.907	0.895	0.749	0.909	0.993	0.912	0.990
	120	0.941	0.910	0.961	0.980	0.953	0.654	0.952	0.990	0.974	0.994

4. สรุปผลและอภิปรายผล

การศึกษาครั้งนี้มีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพของสถิติทดสอบความเท่ากันของเมทริกซ์ความแปรปรวนร่วมของสองประชากรที่มีการแจกแจงปรกติหลายตัวแปรและเป็นข้อมูลที่มีมิติสูง โดยศึกษาสถิติทดสอบ 2 วิธี คือ สถิติทดสอบ LC และสถิติทดสอบ SE โดยมีความผันแปรของเมทริกซ์ความแปรปรวนร่วมตามโครงสร้างของเมทริกซ์ความแปรปรวนร่วมของประชากรเป็น 5 แบบ จากผลการวิจัยสามารถสรุปได้ว่าสถิติทดสอบ LC มีประสิทธิภาพสูงกว่าสถิติทดสอบ SE เนื่องจากสามารถควบคุมความน่าจะเป็นของความคลาดเคลื่อนประเภท

ที่ 1 ได้ทุกโครงสร้างของเมทริกซ์ความแปรปรวนร่วม และมีค่าเข้าสู่ระดับนัยสำคัญ 0.05 เมื่อ n_1, n_2 และ p มีค่าเพิ่มขึ้น โดยโครงสร้าง compound symme-try โครงสร้าง Toeplitz โครงสร้าง unstructured และ โครงสร้าง variance components สถิติทดสอบ LC มีค่าความน่าจะเป็นของความคลาดเคลื่อนประเภทที่ 1 อยู่ในช่วง [0.040,0.060] ถึง 100 % และโครงสร้าง simple มีค่าความน่าจะเป็นของความคลาดเคลื่อนประเภทที่ 1 อยู่ในช่วง [0.040,0.060] ถึง 77 % เมื่อเทียบกับจำนวนผลของค่าความน่าจะเป็นของความคลาดเคลื่อนประเภทที่ 1 ทั้งหมดที่นำเสนอในแต่ละโครงสร้าง และสถิติทดสอบ SE มีค่าความน่าจะเป็นของความคลาดเคลื่อนประเภทที่ 1 ค่อนข้างน้อยในที่นี้เนื่องจากมี n_1, n_2 และ p ที่ยังไม่มากพอ ซึ่งค่าความน่าจะเป็นของความคลาดเคลื่อนประเภทที่ 1 นี้จะดีขึ้นและเข้าสู่ 0.05 เมื่อ n_1, n_2 และ p มีค่ามาก ๆ หรือมีเมตริกซ์ความแปรปรวนร่วมของสองประชากรตาม Srivastava และคณะ [9] ในส่วนของค่าอำนาจการทดสอบพบว่าสถิติทดสอบ LC มีค่าอำนาจการทดสอบสูงกว่าสถิติทดสอบ SE โดยโครงสร้าง Compound symmetry โครงสร้าง simple โครงสร้าง Toeplitz โครงสร้าง unstruck-tured และ โครงสร้าง variance components มีค่าอำนาจการทดสอบสูงกว่าสถิติทดสอบ SE ถึง 100, 77, 88, 66 และ 44 % ตามลำดับ อย่างไรก็ตาม ค่าอำนาจการทดสอบของสถิตทั้งสองต่างยังลู่เข้าสู่ 1 เมื่อ n_1, n_2 และ p มีค่าเพิ่มขึ้น และยิ่งกว่านั้นค่าอำนาจการทดสอบของสถิติทั้งสองมีค่าใกล้เคียงกันมากยิ่งขึ้น เมื่อ n_1, n_2 และ p มีค่าตั้งแต่ 40 ขึ้นไป

5. ข้อเสนอแนะ

5.1 ในงานวิจัยนี้ผู้วิจัยได้ศึกษาเฉพาะกรณีที่ขนาดตัวอย่างเท่ากันทั้งสองกลุ่ม ผู้ที่สนใจอาจศึกษาในกรณีที่ขนาดตัวอย่างไม่เท่ากัน

5.2 ผู้สนใจอาจศึกษาการทดสอบความเท่ากันของเมทริกซ์ความแปรปรวนร่วมของประชากร สำหรับข้อมูลที่มีมิติสูงเมื่อประชากรมีโครงสร้างของเมทริกซ์ความแปรปรวนร่วมอื่น ๆ เช่น โครงสร้าง auto-regressive [10]

5.3 ผู้สนใจเปรียบเทียบสถิติทดสอบอื่น ๆ เช่น สถิติทดสอบของ Chen และ Qin [12] สถิติทดสอบของ Cai และคณะ [13]

6. รายการอ้างอิง

[1] Srivastava, M.S., 2002, Methods of Multivariate Statistics, John Wiley and Sons, Inc., New York, 728p.

[2] กัลยา วานิชย์บัญชา, 2554, การวิเคราะห์ข้อมูลหลายตัวแปร, ภาควิชาสถิติ คณะพานิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย, กรุงเทพ, 589 น.

[3] Anderson, T.W., 2003, An Introduction to Mutivariate Statistical Analysis, 3rd Ed., Wiley-Interscience, New york, pp. 265-267.

[4] Schott, J. R., 2007, A high-dimensional test for the equality of the smallest eigenvalues of a covariance matrix, J. Multivariate Anal. 97: 827-843.

[5] Srivastava, M.S., 2007, Testing the equality of two covariance matrix and testing the independence of two subvectors with fewer observations than the dimension, Proceedings of the International Con-ference on Advanced in Inter-disciplinary Statistics and Combinatorics.

[6] Srivastava, M.S. and Yanagihara, H., 2010,

Testing the equality of several covariance matrices with fewer observations than the dimension, J. Multivariate Anal. 101: 1319-1329.

[7] Li, J. and Chen, S.X., 2012, Two sample test for high dimensional covariance matrices, Anal. Stat. 40: 908-940.

[8] Bai, Z., Jiang, D., Yao, J.F. and Zheng, S., 2009, Corrections to LRT on large-dimensional covariance matrices by RMT, Ann. Statist. 37: 3822-3840.

[9] Srivastava, M.S., Yanagihara, H. and Kunokawa, T., 2014, Test for covariance matrices in high dimension with less sample size, J. Multivariate Anal. 130: 289-309.

[10] Chaipitak, S., 2013, Tests for covariance Matrices with High-Dimensional Data, Ph.D. Thesis, National Institute of Development Administration, Bangkok, 124 p.

[11] Chochran, W.G., 1954, Some methods for strengthening the common λ^2 test, Biometrics 10: 417-451.

[12] Chen, S.X. and Qin, Y.L., 2010, A two-sample test for high-dimensional data with applications to gene-set testing, Anal. Stat. 38: 808-835.

[13] Cai, T., Lui, W. and Xai, Y., 2013 Two-sample covariance matrix testing and support recovery in high-dimensional and sparse settings, J. Amer. Stat. Assoc. 108: 265-277.