Evaluating Test Statistics for Covariance Matrix Equality in Multivariate Repeated-Measures Data: A Comprehensive Simulation Study

***AJIBOYE, Raimot A. and OYEYEMI, Gafa M.**

Department of Statistics, University of Ilorin, Ilorin, Nigeria

*Corresponding email: 01-55EG013pg@students.unilorin.edu.ng, ajiboyera@gmail.com

Abstract

This study evaluates the performance of the Likelihood Ratio Test (LRT), Box's M test, Nagao's Trace test, and Ahmad's Tau statistics for testing the equality of covariance matrices in multivariate repeated-measures data. Using extensive Monte Carlo simulations that vary the number of variables, groups, and sample sizes, the study compares the tests in terms of Type I error control and statistical power. Results show that while LRT provides high power, it exhibits inflated Type I error rates in small samples and higher dimensions. Box's M test consistently maintains appropriate Type I error rates and demonstrates robust power across diverse scenarios, making it a reliable choice for moderate-dimensional data. Nagao's Trace test tends to be conservative with lower power in smaller samples, and Ahmad's Tau statistics are overly conservative, limiting their practical utility. These findings offer guidance for researchers in biomedical, psychological, and social sciences in selecting suitable methods for covariance matrix equality testing, emphasizing a balance between Type I error control and statistical power.

Keywords: Covariance matrix equality; Type I error control; Likelihood Ratio Test; Box's M test; Nagao test; multivariate repeated measures

1.0 Introduction

The equality of variance-covariance matrices across groups is a fundamental assumption in many multivariate statistical procedures, especially in repeated-measures and longitudinal designs. In these settings, the same subjects are measured multiple times under varying conditions, producing correlated multivariate data. Accounting for intra-subject correlation reduces variability due to individual differences and increases statistical efficiency (Field, Miles, & Field, 2012). Accurate estimation and comparison of covariance matrices are therefore crucial for hypothesis testing and model validation.

Repeated-measures designs have been widely studied under different terminologies, including within-subjects ANOVA, treatments-by-subjects ANOVA, and randomized-blocks ANOVA (Maxwell, Delaney, & Kelley, 2017). Extensions to two-way repeated-measures allow the assessment of interaction effects across multiple factors over time (Keselman, Algina, & Kowalchuk, 2001). Central to these analyses is the assumption of homogeneous covariance structures across groups or conditions; violations can lead to biased inferences, highlighting the need for robust statistical tools to test covariance equality.

Classical approaches include Box's M test (Box, 1949, 1950), which compares pooled and group-specific covariance matrices under multivariate normality. While widely used, Box's M is sensitive to deviations from normality and small sample sizes (O'Brien, 1992; Muirhead, 2009). Nagao (1973) proposed a trace-based statistic that evaluates covariance equality using the squared deviations of individual covariance matrices from the pooled matrix. More recently, Ahmad (2021) introduced three Tau statistics for two-group comparisons, designed to provide robust alternatives under moderate sample sizes.

Modern studies have addressed limitations of classical tests under non-normal distributions and high-dimensional data (Srivastava & Yanagihara, 2010; Cai, Liu, & Xia, 2013; Zi & Chen, 2022; Wang, Liu, & Zhang, 2024). Robust methods for covariance estimation in the presence of outliers and heavy-tailed distributions have also been developed (Oyeyemi & Ipinyomi, 2010; Vandev & Neykov, 2000). Recent work (Zhang et al., 2021; Liu & Chen, 2022; Gupta et al., 2023) has proposed modified test statistics that improve performance in small-to-moderate sample sizes and under non-normality, highlighting ongoing advances in practical multivariate analysis.

Despite these advances, most high-dimensional methods assume $p > n$ and often rely on restrictive assumptions, limiting applicability to moderate-dimensional repeated-measures data where $p < n$ but classical assumptions may still fail. Schott (2001) evaluated Wald-type test statistics under broader elliptical distributions and demonstrated that the traditional LRT is liberal under positive kurtosis and conservative under negative kurtosis. He proposed a generalized Wald test with improved reliability across moderate-dimensional, non-normal contexts.

This study focuses on evaluating the performance of classical and modern test statistics—Box's M, LRT, Nagao, and Ahmad's Tau tests—in scenarios where the number of variables is moderate relative to the sample size. The objective is to assess Type I error control and statistical power under varying group numbers, sample sizes, and covariance structures. By empirically investigating these conditions, the study provides guidance on the practical applicability of existing methods and identifies their limitations in moderate-dimensional repeated-measures designs.

2.0 Methodology

This section outlines both the theoretical framework and simulation procedures used to evaluate the performance of various test statistics for assessing the equality of covariance matrices. The statistical methods reviewed include the Likelihood Ratio Test (LRT), Box's M Test, Nagao's Trace Test, and three competing statistics introduced by Ahmad (2021).

2.1 Test Statistics Framework

Let $\mathbf{X}_i = \left(\mathbf{X}_{i1}^\top, \mathbf{X}_{i2}^\top, \dots, \mathbf{X}_{in_i}^\top\right)^\top$ denote an $n_i \times p$ matrix of observations from the i-th group, drawn from a p-variate normal distribution $\mathcal{N}_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$. Define $n = \sum_{i=1}^{g} n_i$ as the total sample size.

The sample mean and sample covariance matrix for group i are given by:

$$\bar{\mathbf{X}}_i = \frac{1}{n_i}\sum_{k=1}^{n_i} \mathbf{X}_{ik}, \quad \hat{\boldsymbol{\Sigma}}_i = \frac{1}{n_i - 1}\sum_{k=1}^{n_i} (\mathbf{X}_{ik} - \bar{\mathbf{X}}_i)\,(\mathbf{X}_{ik} - \bar{\mathbf{X}}_i)^\top \tag{2.1}$$

To test the hypothesis:

$$H_0\colon \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \cdots = \boldsymbol{\Sigma}_g = \boldsymbol{\Sigma}, \tag{2.2}$$

Box (1949, 1950) proposed the following LRT-based test statistic:

$$C = (1-u)M, \tag{2.3}$$

where:

$$\begin{aligned} M &= (n-g)\ln\left|\widehat{\boldsymbol{\Sigma}}\right| - \sum_{i=1}^{g}(n_i-1)\ln\left|\widehat{\boldsymbol{\Sigma}}_i\right|, \\ a &= \sum_{i=1}^{g}\left(\frac{1}{n_i-1}\right) - \frac{1}{n-g}, \quad b = \frac{2p^2+3p-1}{6(p+1)(g-1)}, \quad u = ab. \end{aligned} \tag{2.4}$$

The pooled covariance matrix is defined as:

$$\widehat{\boldsymbol{\Sigma}} = \frac{1}{n-g}\sum_{i=1}^{g}(n_i-1)\,\widehat{\boldsymbol{\Sigma}}_i \tag{2.5}$$

Under multivariate normality, C follows a chi-square distribution asymptotically with degrees of freedom:

$$f = \frac{1}{2}p(p+1)(g-1) \tag{2.6}$$

Nagao (1973) proposed a trace-based alternative:

$$N = \frac{1}{2}\sum_{i=1}^{g} n_i \cdot \operatorname{tr}\left[\left(\widehat{\boldsymbol{\Sigma}}_i\widehat{\boldsymbol{\Sigma}}^{-1} - \mathbf{I}\right)^2\right], \tag{2.7}$$

where **I** is the identity matrix.

Ahmad (2021) introduced three alternative test statistics for the two-group case ($g = 2$):

$$\begin{aligned} \tau_1 &= \operatorname{tr}\left(\widehat{\boldsymbol{\Sigma}}_1\widehat{\boldsymbol{\Sigma}}_2^{-1}\right), \\ \tau_2 &= \operatorname{tr}\left[\widehat{\boldsymbol{\Sigma}}_1\left(\widehat{\boldsymbol{\Sigma}}_1 + \widehat{\boldsymbol{\Sigma}}_2\right)^{-1}\right], \\ \tau_3 &= \frac{\left|\widehat{\boldsymbol{\Sigma}}_1 + \widehat{\boldsymbol{\Sigma}}_2\right|}{\left|\widehat{\boldsymbol{\Sigma}}_2\right|}. \end{aligned} \tag{2.8}$$

2.2 Data-Generating Procedure

A simulation study was conducted to empirically assess the performance of the test statistics under varying conditions.

1. **Type I Error Evaluation:** Data for all groups were generated from the same multivariate normal distribution $Np(\mu, \Sigma)$ representing the null hypothesis.
2. **Power Evaluation:** Data for each group were generated from distinct multivariate normal distributions $Np(\mu, \Sigma_i)$ with different covariance matrices, representing the alternative hypothesis.
3. **Covariance Matrices:** Pre-specified population covariance matrices were used to provide controlled scenarios for evaluating both Type I error and statistical power under known structures.
4. **Iteration:** The simulations were repeated k times to ensure stability of results. Empirical Type I error and power were computed as the proportion of hypothesis rejections across iterations.
5. **R Implementation:** Seed initialization (set.seed(42)) and mvrnorm() from the MASS package are used for reproducibility.

3.0 Discussion of Results

This section presents a detailed evaluation of the performance of various statistical tests used for assessing the equality of covariance matrices across multiple groups. The evaluation focuses on Type I error rates and power under different combinations of dimensionality (p) and number of groups (g) across varying sample sizes. The results are interpreted in the context of practical applicability and robustness of the Likelihood Ratio Test (LRT), Box's M test, Nagao test, and Ahmad's Tau tests.

Table 1: Comparison of Type I error rates for LRT, Box's M, Nagao, and Tau tests at varying sample sizes.

Sample Size	LRT	Box's M	Nagao	Tau1	Tau2
10	0.120	0.047	0.030	0.207	0.380
20	0.071	0.053	0.022	0.027	0.178
30	0.073	0.058	0.026	0.005	0.077
50	0.057	0.053	0.029	0.001	0.026
80	0.064	0.061	0.031	0.000	0.009
100	0.055	0.051	0.030	0.000	0.004
200	0.056	0.053	0.027	0.000	0.000
300	0.046	0.046	0.026	0.000	0.000
500	0.040	0.039	0.026	0.000	0.000

At small sample sizes, the Likelihood Ratio Test (LRT) tends to be slightly liberal, exceeding the nominal 0.05 Type I error level, whereas Box's M and Nagao tests remain closer to the nominal level, indicating better control of false positives, as shown in Table 1. Tau1 and Tau2 tests are overly conservative or unstable at small sample sizes, with very low or zero Type I errors, indicating potential under-rejection when the null hypothesis is true. As sample size increases, LRT, Box's M, and Nagao tests show improved adherence to the nominal level, confirming their suitability for moderate to large sample sizes.

Table 2: Empirical power of LRT, Box's M, Nagao, and Tau tests across different sample sizes.

Sample Size	LRT	Box's M	Nagao	Tau1	Tau2
10	0.171	0.069	0.030	0.074	0.144
20	0.194	0.129	0.059	0.001	0.015
30	0.297	0.260	0.159	0.000	0.001
50	0.480	0.455	0.348	0.000	0.001
80	0.708	0.692	0.622	0.000	0.000
100	0.821	0.808	0.760	0.000	0.000
200	0.991	0.991	0.987	0.000	0.000
300	0.998	0.998	0.998	0.000	0.000
500	1.000	1.000	1.000	0.000	0.000

The power results in Table 2 show that LRT, Box's M, and Nagao tests substantially increase their ability to detect true differences in covariance matrices as the sample size grows. Nagao test tends to have slightly lower power compared to LRT and Box's M at small sizes, but catches up at larger sizes. Tau1 and Tau2 tests demonstrate negligible power, reinforcing the earlier observation that these tests may be too conservative or unsuitable for moderate sample sizes and dimension configurations.

Table 3: Type I error rates of LRT, Box's M, and Nagao tests for three variables and three groups.

Sample Size	LRT	Box's M	Nagao
10	0.315	0.057	0.057
20	0.125	0.058	0.015
30	0.107	0.064	0.017
50	0.069	0.055	0.014
80	0.059	0.053	0.012
100	0.057	0.052	0.013
200	0.055	0.053	0.010
300	0.066	0.062	0.014
500	0.049	0.047	0.011

For $p = 3$ and $g = 3$ in Table 3, LRT exhibits substantial Type I error inflation at very small sample sizes, far exceeding the nominal level. Box's M and Nagao tests maintain better control, with error rates closer to the nominal 0.05 level across all sample sizes. This suggests that for higher dimensions or more groups, Box's M and Nagao are more reliable for Type I error control in small to moderate sample sizes.

Table 4: Power comparison of LRT, Box's M, and Nagao tests for p = 3 and g = 3.

Sample Size	LRT	Box's M	Nagao
10	0.435	0.095	0.058
20	0.516	0.334	0.088
30	0.746	0.645	0.380
50	0.957	0.943	0.854
80	0.997	0.997	0.991
100	1.000	1.000	1.000
200	1.000	1.000	1.000
300	1.000	1.000	1.000

Power improves significantly for all tests as the sample size increases, as shown in Table 4. LRT generally outperforms Box's M and Nagao in power at small to moderate sizes, but suffers from inflated Type I errors, as seen earlier. Nagao test remains more conservative but gains power rapidly with increasing sample size, balancing power and error control effectively.

Table 5: Type I error rates for LRT, Box's M, and Nagao tests with four variables and four groups.

Sample Size	LRT	Box's M	Nagao
10	0.712	0.163	0.075
20	0.195	0.064	0.009
30	0.116	0.050	0.002
50	0.079	0.047	0.003
80	0.076	0.059	0.005
100	0.066	0.053	0.004
200	0.062	0.054	0.003
300	0.077	0.069	0.003
500	0.050	0.047	0.004

Table 5 shows that as dimensionality increases to 4 variables and 4 groups, LRT shows extreme Type I error inflation at small sample sizes, which reduces but remains above nominal levels. Box's M and Nagao tests remain stable but slightly conservative. This indicates that LRT assumptions break down in higher dimensions and small samples, making Box's M and Nagao preferable.

Table 6: Power of LRT, Box's M, and Nagao tests for four variables and four groups.

Sample Size	LRT	Box's M	Nagao
10	0.817	0.170	0.078
20	0.830	0.563	0.117
30	0.964	0.911	0.577
50	0.998	0.997	0.983
80	1.000	1.000	1.000
100	1.000	1.000	1.000
200	1.000	1.000	1.000
300	1.000	1.000	1.000
500	1.000	1.000	1.000

Power is high for all tests as sample sizes increase, with LRT maintaining the highest power but at the cost of inflated Type I error in Table 6. Box's M provides a good balance between power and error control. Nagao test lags in power at small sample sizes but rapidly approaches full power at moderate sizes.

Table 7: Type I error rates for LRT, Box's M, and Nagao tests with five variables and five groups.

Sample Size	LRT	Box's M	Nagao
10	NaN	NaN	NaN
20	0.301	0.072	0.011
30	0.148	0.049	0.002
50	0.113	0.054	0.001
80	0.076	0.049	0.001
100	0.071	0.046	0.000
200	0.065	0.056	0.002
300	0.041	0.034	0.000
500	0.058	0.053	0.002

Table 7 provides that at five variables and five groups, LRT has unstable results at very small sample sizes (NaNs), and inflated Type I error at low sample sizes. Box's M and Nagao maintain well-controlled error rates throughout, affirming their robustness in complex settings.

Table 8: Power of LRT, Box’s M, and Nagao tests for five variables and five groups.

Sample Size	LRT	Box’s M	Nagao
10	NaN	NaN	NaN
20	0.969	0.694	0.083
30	0.996	0.979	0.698
50	1.000	1.000	1.000
80	1.000	1.000	1.000
100	1.000	1.000	1.000
200	1.000	1.000	1.000
300	1.000	1.000	1.000
500	1.000	1.000	1.000

Despite unstable LRT performance at very small samples, all tests achieve perfect power at moderate to large sample sizes, confirming that they effectively detect true covariance differences when sample size permits, as shown in Table 8.

Conclusion

This study evaluated the performance of several test statistics for assessing the equality of covariance matrices in multivariate repeated-measures settings, focusing on scenarios with moderate numbers of variables relative to sample size. Extensive simulations examined the Likelihood Ratio Test (LRT), Box’s M test, Nagao test, and Ahmad’s Tau statistics in terms of Type I error control and statistical power across varying dimensions, numbers of groups, and sample sizes.

The results show that Box’s M test consistently maintains appropriate Type I error rates and demonstrates strong power across a wide range of conditions, particularly with moderate to large sample sizes. In contrast, LRT exhibits inflated Type I errors in small samples and higher dimensions. Nagao’s test tends to be conservative with lower power in small samples, and Ahmad’s Tau statistics are often overly conservative, failing to reject the null hypothesis under alternative scenarios.

Overall, Box’s M test is highlighted as a robust and reliable method for testing covariance matrix equality in moderate-dimensional multivariate data, particularly in repeated-measures designs commonly encountered in biomedical, psychological, and social science research. These findings provide practical guidance for selecting appropriate tests in studies where classical assumptions may be challenged.

Competing Interests

The authors declare that they have no known competing financial or non-financial interests that could have appeared to influence the work reported in this paper. There are no conflicts of interest related to the design, execution, or interpretation of the research presented.

Declaration of Funding
This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Ethics Approval

This study does not involve human participants or animals and therefore does not require ethics approval.

Consent to Participate

Not applicable.

Consent for Publication

All authors have given their consent for publication of this manuscript.

Availability of Data and Materials

The datasets analyzed during the current study were synthetically generated using multivariate normal distributions, as detailed in the Methodology section. These datasets were created under both the null and alternative hypotheses to evaluate the performance of various test statistics for testing the equality of covariance matrices. All data generation followed a reproducible design, with fixed random seed initialization. The code and scripts used for data generation, test statistic computation (including Box's M Test, Likelihood Ratio Test, Nagao's Trace Test, and Ahmad's τ statistics), and simulation analysis are available from the corresponding author upon reasonable request.

Authors' contributions:
R.A.A.: Conceptualization, Methodology, Resources, Formal Analysis, Writing Original Draft, Data Curation

G.M.O.: Resources, Review, and Editing

References

Ahmad, R. (2021). A robustness evaluation of homogeneity test of covariance matrices. In *Proceedings of the Fifteenth International Conference on Management Science and Engineering Management: Volume 1* (pp. 309–321). Springer International Publishing.

Box, G. E. (1949). A general distribution theory for a class of likelihood criteria. *Biometrika*, 36(3/4), 317–346.

Box, G. E. (1953). Non-normality and tests on variances. *Biometrika*, 40(3/4), 318–335.

Bulut, H. (2025). A Novel Robust Test to Compare Covariance Matrices in High-Dimensional Data. *Axioms*, *14*(6), 427.

Cai, T., Liu, W., & Xia, Y. (2013). Two-sample covariance matrix testing and support recovery in high-dimensional and sparse settings. *Journal of the American Statistical Association*, 108(501), 265–277.

Cortez-Elizalde, D., & Bolivar-Cime, A. (2022). Behavior of Some Hypothesis Tests for the Covariance Matrix of High Dimensional Data. *Revista Colombiana de Estadística*, *45*(2), 373-399.

Ding, X., Hu, Y., & Wang, Z. (2024). Two sample test for covariance matrices in ultra-high dimension. *Journal of the American Statistical Association*, 1-12.

Field, A., Field, Z., & Miles, J. (2012). Discovering statistics using R.

Jiang, Y., Wen, C., Jiang, Y., Wang, X., & Zhang, H. (2023). Use of random integration to test equality of high dimensional covariance matrices. *Statistica Sinica*, *33*(4), 2359.

Keselman, H. J., Algina, J., & Kowalchuk, R. K. (2001). The analysis of repeated measures designs: a review. *British Journal of Mathematical and Statistical Psychology*, *54*(1), 1-20.

Maxwell, S. E., Delaney, H. D., & Kelley, K. (2017). *Designing experiments and analyzing data: A model comparison perspective*. Routledge.

Muirhead, R. J. (2009). *Aspects of multivariate statistical theory*. Wiley.

Nagao, H. (1973). On some test criteria for covariance matrix. *The Annals of Statistics*, 700–709.

O’Brien, P. C. (1992). Robust procedures for testing equality of covariance matrices. *Biometrics*, 819–827.

Oyeyemi, G. M., & Ipinyomi, R. A. (2010). A robust method of estimating covariance matrix in multivariate data analysis. *African Journal of Mathematics and Computer Science Research*, 3(1), 1–18.

Schott, J. R. (2001). Some tests for the equality of covariance matrices. *Journal of Statistical Planning and Inference*, 94(1), 25–36.

Singh, V., Rana, R. K., & Singhal, R. (2013). Analysis of repeated measurement data in the clinical trials. *Journal of Ayurveda and Integrative Medicine*, 4(2), 77.

Zi, X., & Chen, H. (2022). Robust tests of the equality of two high-dimensional covariance matrices. Communications in Statistics-Theory and Methods, 51(10), 3120-3141.

Srivastava, M. S., & Yanagihara, H. (2010). Testing the equality of several covariance matrices with fewer observations than the dimension. *Journal of Multivariate Analysis*, 101(6), 1319–1329.

Vandev, D. L., & Neykov, N. M. (2000). Robust maximum likelihood in the Gaussian case. Retrieved from http://www.fmi.uni-sofia.bg/fmi/statist/Personal/Vandev/papers/ascona_1992.pdf

Wang, J., Zhu, T., & Zhang, J. T. (2024). Two-sample test for high-dimensional covariance matrices: A normal-reference approach. *Journal of Multivariate Analysis*, 204, 105354.

Wang, J., Zhu, T., & Zhang, J. T. (2025). Test of the Equality of Several High-Dimensional Covariance Matrices: A Normal-Reference Approach. *Mathematics*, *13*(2), 295.

Yu, L., Xie, J., & Zhou, W. (2023). Testing Kronecker product covariance matrices for high-dimensional matrix-variate data. *Biometrika*, *110*(3), 799-814.