ANOVA on principal component as an alternative to MANOVA

Bahriddin Abapihi\(^1\), Gusti Ngurah Adhi Wibawa\(^1\), Baharuddin\(^1\), Mukhsar\(^1\)*, Agusrawati\(^1\) and Lillis LaOme\(^1\)

\(^1\)Department of Statistics, Halu Oleo University, Kendari, Indonesia

mukhsarmuhu@gmail.com (corresponding author)

Abstract. With its strict assumptions, practitioners found difficulties in applying Multivariate Analysis of Variance (MANOVA) on their works. When normal assumption is partially fulfilled on multivariate responses, it does not guarantee that the responses are to be multivariate normal distributed simultaneously. To tackle this problem, we proposed a method by simply applying Analysis of Variance (ANOVA) on principal component (PC). The PC is a linear combination of the responses. Once all the responses are normally distributed, the PCS will also be. Accordingly, all we need is to ensure that the responses are to be partially normal distributed and the multivariate normal distribution is needless. The results of our data analysis indicated that our proposed method can be used as an alternative to MANOVA, especially when multivariate normal assumption could not be fully guaranteed.

1. Introduction

Multivariate Analysis of Variance (MANOVA) in experimental design is employed to investigate the effects of treatments on multi-responses simultaneously instead of using Analysis of Variance (ANOVA) partially [1], [2], [6]. However, practitioners found difficulties in applying MANOVA on their works due to the underlying assumptions to properly using MANOVA. One of the underlying assumptions is the multivariate normality which is difficult to fulfill. When assumption of normality is fulfilled on responses separately, it could not be assured that the responses are to be simultaneously multivariate normal distributed. This strict assumption is practically difficult to be satisfied. Accordingly, the application of MANOVA to such a case is no longer suitable.

Like ANOVA, a MANOVA mainly aims to test the significance of treatments and the interaction between treatments (if the model containing interaction term). When the null hypothesis (that all treatments are the same) is rejected, further analysis needs to be performed to draw conclusion about which treatment measures differ from others. Interpretation on interaction term is also another problem to deal with when it is significant. To overcome the problems above, we proposed an application of ANOVA on principal component (PC) instead of applying MANOVA on multivariate responses. This approach takes consideration since principal component is a linear combination of original variables, the responses. And as we know, when response variables are separately having normal distribution, the principal components will also be. By ensuring each of response variables to be normally distributed, it will be confirmed that the principal components are also to follow normal distribution. With PCS, we only apply ANOVA on a single response variable, i.e. the principal component score which has the highest contribution to the total variation of original response variables.
2. Principal Component Analysis (PCA)

The PCA is a technique to transform a set of $p$ original variables, usually correlated, say $X_1, X_2, \ldots, X_p$, to a new set of $r$ ($r \leq p$) uncorrelated variables, say $Y_1, Y_2, \ldots, Y_r$. These new set of variables are called principal components, the linear combinations of original variables. Geometrically, these linear combinations represent the new coordinate system by rotating the original coordinate system with $Y_1, Y_2, \ldots, Y_r$ as the axes. These axes represent the directions with maximum variability on covariance structure [1], [3], [4], [5].

Suppose random vector $X' = [X_1, X_2, \ldots, X_p]$ having covariance matrix $\Sigma$ with corresponding characteristic roots $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$. Consider the following linear combinations.

\[
Y_1 = a_1'X = a_{11}X_1 + a_{12}X_2 + \ldots + a_{1p}X_p \\
Y_2 = a_2'X = a_{21}X_1 + a_{22}X_2 + \ldots + a_{2p}X_p \\
\vdots \\
Y_p = a_p'X = a_{p1}X_1 + a_{p2}X_2 + \ldots + a_{pp}X_p
\]  

We get $\text{Var}(Y_i) = a_{ii}'\Sigma a_i$ $i = 1, 2, \ldots, p$ and $\text{Cov}(Y_i, Y_j) = a_{ij}'\Sigma a_j$ $i, j = 1, 2, \ldots, p$. $Y_1, Y_2, \ldots, Y_p$ are principal components, the linear combinations of original variables which are uncorrelated among others. The first principal component explaining most of total variance can be obtained by maximizing $\text{Var}(Y_1) = a_{11}'\Sigma a_1$ subject to $a_{11}'a_1 = 1$. For the second principal component is to maximize $\text{Var}(Y_2) = a_{22}'\Sigma a_2$ with constraints $a_{22}'a_2 = 1$ and $\text{Cov}(Y_2, Y_1) = 0$. And so forth for $i$th principal component by assigning $\text{Var}(Y_i) = a_{ii}'\Sigma a_2$ subject to $a_{ii}'a_i = 1$ and $\text{Cov}(Y_i, Y_j) = 0$ for any $j < i$.

3. ANOVA and MANOVA Model

MANOVA model is similar to ANOVA. The only difference is the use of multivariate response instead of single variable separately. In this paper we only show a two-factor model with interaction. For a two-factor ANOVA, the model with interaction can be written as:

\[
x_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk}
\]  

Where $\Sigma \alpha_i = \Sigma \beta_j = \Sigma \gamma_{ij} = 0$ and $\epsilon_{ijk}$ be normally independent distributed with zero mean and variance $\sigma^2$. In this model, $\mu$ represents general mean, $\alpha_i$ be fixed effect of factor I, $\beta_j$ be fixed effect of factor II, and $\gamma_{ij}$ be the interaction between factor I and factor II [1], [3]. Like in ANOVA model, MANOVA model can be written as Equation (2) with all terms are vectors with response term in matrix, as follows.

\[
X_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk}
\]  

Where $\Sigma \alpha_i = \Sigma \beta_j = \Sigma \gamma_{ij} = 0$. All vectors are of order $p \times 1$ ($p$ is the number of response variables) and $\epsilon_{ijk}$ are random vectors of independent multivariate normal $N_p(0, \Sigma)$ [1].

4. Results and Discussions

To show that our proposed method work well, we utilized an artificial dataset with three response variables, two factors with 4 levels for each, and the number of replication is 5. The first step of this section we would be providing an ANOVA to see if the factors and interaction are showing significant effects on separated response variable. Then, we would be presenting the results from MANOVA to compare the effects on response variables simultaneously. Finally we would be performing the results
of ANOVA on principal component (PC ANOVA) as a single response. Note that if the factor and interaction effects in all methods (at least for MANOVA and PC ANOVA) showing the same level of significance, then it means that the results are confirming each other. Consequently, we could expect that PC ANOVA is applicable to conduct multivariate sample means comparison, especially when multivariate normal assumption could not be fulfilled.

4.1. ANOVA on separated response variables
Analysis of variance separately on responses are presented in Table 1. As shown in the table, Factor A is only significant on the response variable $X_1$, while Factor B and the interaction effects are significant on all response variables. The replication effects are not significant on all responses. Note that in all tables df stands for degree of freedom. The critical value is $\alpha = 0.05$, which means that null hypothesis is rejected when $p$-value less than $\alpha$.

| Source of Variation | df | $X_1$ | $X_2$ | $X_3$ |
|---------------------|----|-------|-------|-------|
| Factor A            | 3  | 0.000 | 0.820 | 0.262 |
| Factor B            | 3  | 0.000 | 0.000 | 0.000 |
| Interaction A*B     | 9  | 0.000 | 0.000 | 0.000 |
| Replication         | 4  | 0.883 | 0.343 | 0.563 |
| Error               | 60 |       |       |       |
| Total               | 79 |       |       |       |

4.2. MANOVA on all response variables simultaneously
The application of MANOVA on all response variables simultaneously shows that all factors and the interaction are to have significant effects on the responses (Table 2). The effects of replication simultaneously on responses are not significant like using ANOVA on separated response variables.

| Source of Variation | Df       | Test Criteria | p-values |
|---------------------|----------|---------------|----------|
| Factor A            | (9,141)  | Wilk’s        | 0.000    |
|                     | (9,170)  | Lawley-Hotelling | 0.000    |
|                     | (9,180)  | Pillai’s      | 0.000    |
| Factor B            | (9,141)  | Wilk’s        | 0.000    |
|                     | (9,170)  | Lawley-Hotelling | 0.000    |
|                     | (9,180)  | Pillai’s      | 0.000    |
| Interaction A*B     | (27,170) | Wilk’s        | 0.000    |
|                     | (27,170) | Lawley-Hotelling | 0.000    |
| Replication         | (27,180) | Pillai’s      | 0.000    |
|                     | (12,153) | Wilk’s        | 0.758    |
|                     | (12,170) | Lawley-Hotelling | 0.770    |
|                     | (12,180) | Pillai’s      | 0.746    |
4.3. ANOVA on principal component

The ANOVA is conducted only on the first principal component (PC). Based on the PCA result, the first PC explains 53% of total variance, a reliable amount of variance even though it is not good enough. Using the scores of first PC, the ANOVA took place. The results can be seen on Table 3.

| Source of Variation | df | p-values |
|---------------------|----|----------|
| Factor A            | 3  | 0.000    |
| Factor B            | 3  | 0.000    |
| Interaction A*B     | 9  | 0.000    |
| Replication         | 4  | 0.767    |
| Error               | 60 |          |
| Total               | 79 |          |

We can see from Table 3 that all factors and interaction, except replication, are significant. If we compare the results of using MANOVA as presented in Table 2 and of ANOVA on principal component as in Table 3, both methods are confirming each other. The levels of significance of corresponding factors in either table are almost the same. This is an indication that both methods producing the same results. We can choose whether to use one method or the other. But, for sure, using ANOVA on principal component is more convenient since no strict multivariate normal assumption is required.

5. Conclusion

We have shown that the use of MANOVA on multivariate responses is producing the similar result as of ANOVA on a single principal component. Of course, using the last one is more effective and easy to interpret instead, since no need to fulfil the multivariate normal assumption.

References

[1]   Hedeker, D. & Gibbons, R. D. 2006. Longitudinal Data Analysis. New Jersey: Wiley.
[2]   Zhang, J. T. & Xiao, S. 2012. A note on the modified two-way MANOVA tests. Statistics and Probability Letters 82 (2012) 519-527.
[3]   Johnson, R. A. & Wichern, D. W. 1998. Applied Multivariate Statistical Analysis. London: Prentice-Hall.
[4]   Gomez, K. A. & Gomez, A. A. 1984. Statistical Procedure for Agricultural Research. New Jersey: Wiley.
[5]   Morrison, D. F. 1976. Multivariate Statistical Methods. New York: McGraw-Hill.
[6]   Wang, M., Zhou, Y., & Tan, G. Z. 2019. Multivariate analysis of variance (MANOVA) on the microstructure gradient of biomimetic nanofiber scaffolds fabricated by cone electrospinning. Journal of Manufacturing Processes 44 (2019) 55-61.