https://doi.org/10.4038/sljastats.v24i2.8094

Sri Lankan Journal of Applied Statistics, Vol (24-2)

Incorporation of Covariates Through Principal Components in Analysis of Covariance: A Simulation Study

A. S. G. Jayasinghe[1,*], **and S. Samita**[2]

[1]Postgraduate Institute of Agriculture, University of Peradeniya, Sri Lanka

[2]Faculty of Agriculture, University of Peradeniya, Sri Lanka

*Corresponding author: gayathreejayasinghe92@gmail.com

Received: 26th June 2023/ Revised: 29th August 2023/ Published: 30th September 2023

ABSTRACT

This study examined the use of the Principal Component Analysis (PCA) approach in incorporating covariates in the Analysis of Covariance (ANCOVA) under different experimental setups. Simulated data were used for the study, and the statistical programs were developed in R statistical software to generate the datasets for different experimental setups by varying (i) number of covariates, (ii) degree of correlation among covariates, (iii) number of treatments, (iv) difference between treatment means, and (v) number of replicates. Thousand simulations were performed for each experimental setup, and the impact of the PCA approach was assessed by means of power of the test through the proportion of rejections of H0: no difference between adjusted treatment means, in 1000 simulations. The use of PCs led to a significant gain in the power of the test in ANCOVA when there is a higher number of interrelated covariates with a limited number of observations. The impact was higher with the increase of number of covariates as well as the correlation between covariates. It can be concluded that by accommodating covariates by means of PCs, the efficiency in ANCOVA can be increased, especially if there are many covariates to be included in the analysis with a limited number of observations.

Keywords: Power of the test, Degree of correlation, Dimensionality reduction, Interrelated covariates, Treatment effect.

1 Introduction

Analysis of Covariance (ANCOVA) is an extension of analysis of variance (ANOVA), and it is used to increase the precision in testing the null hypothesis that two or more population means are equal (Huitema 2011). R. A. Fisher originally developed this statistical technique in 1932, to reduce error variance in experimental studies (Shieh 2020). Since then, the technique has been further developed and expanded for application in agriculture and other disciplines (Yang and Juskiw 2011). ANCOVA techniques are often employed in the analysis of clinical trials to try to account for the effects of varying pretreatment baseline values of an outcome variable on post treatment measurements of the same variable (Crager 1987).

ANCOVA provides a way of statistically controlling the (linear) effect of variables, which does not want to examine in a study, and typically used to adjust or control for differences between the groups, based on another. ANCOVA always involves at least three variables: an independent variable, a dependent variable, and a covariate (Huitema, 2011). Studying the effect of covariates is not usually the primary interest of ANCOVA, but they are typically used to increase the precision of estimates of other source variables through control of experimental error (Huitema, 2011). Often an experiment has one or two covariates, selected before analyzing the data. However, there can be a substantial number of covariates to include in the analysis to make the analysis more precise, where the number of observations is limited. In such a case, only one or two covariates have to be selected to preserve the error degree of freedom (DF). Raab et al. (2000) stated that there is a wide range of variable selection methods, such as forward or backward stepwise procedures, which can be used to decide which set of covariates to include in the analysis. Nevertheless, improving analysis by ignoring covariates seems counterintuitive, as they could provide some information (Mefford and Witte, 2012). Therefore, finding a way to use all the possible covariates without loss of error DF effectively, would be the most promising solution. A possible approach could be to incorporate covariates by means of one or few composite variables where a composite variable is a variable made up of two or more variables or measures that are highly related to one another conceptually or statistically (Song et al. 2013). The composite variable is then can be used as a covariate in the main analysis while minimizing the loss of DF.

The use of composites in analyses has been discussed in many studies (Clements et al., 2022; Branders et al., 2021) with different methods of constructing composites. The principal component analysis (PCA) approach is one of the possible methods, which can be used in constructing composite variables. Although there were studies on the use of PCA approach in constructing composite indices (Senna et al. 2019; Dharmawardena et al., 2016; Li et al., 2012), there were limited studies on the use of the PCA approach in constructing composite covariates in ANCOVA.

The central idea of PCA is to reduce the dimensionality of a dataset consisting of a large number of interrelated variables, while retaining as much as possible of the variation present in the data set. This is achieved by transforming to a new set of variables, the principal components (PCs), which are uncorrelated, and ordered, so that the first few retain most of the variation present in all of the original variables (Jolliffe 2002). Since the PCA approach reduces the dimensionality of a dataset, this can be used to reduce the number of covariates, by replacing the covariates with PCs. Therefore, this study aimed to investigate the use of the PCA approach in incorporating covariates, by means of PCs in the ANCOVA when there is substantial number of covariates.

2 Methodology

Simulated data were used for the study and the R software package (Version - R 4.1.3) was used to simulate the data. The statistical programs were developed to generate simulated datasets involving specific experimental setups, and analysis. The covariates were considered as continuous variables and generated under the multivariate normal distribution with random errors (ε) satisfying independent normal distributions. The simulation model which used to generate data sets is given as equation 1.

$$y_{ij} = \mu + \tau_i + \beta_1(x_{ij1} - \bar{x}_{..1}) + \beta_2(x_{ij2} - \bar{x}_{..2}) + ... + \beta_p(x_{ijp} - \bar{x}_{..p}) + \epsilon_{ij}$$

$$i = 1, 2, ..., t, j = 1, 2, ..., n_i \tag{1}$$

where, μ = Grand mean, τ_i= i^{th} Treatment effect, $\beta_1, \beta_2, ..\beta_p$ = Regression coefficients, $x_{ij1}, x_{ij2}, ..., x_{ijp}$ are value of the covariate of j^{th} observation of i^{th} treatment, $\bar{x}_{..1}, \bar{x}_{..2}, ..., \bar{x}_{..p}$ are means of the respective covariates and ϵ_{ij} = random errors.

In this study, different setups were obtained by varying (i) number of covariates, (ii) degree of correlation among covariates, (iii) number of treatments, (iv) difference between treatments, and (v) number of replicates. For each experimental setup, 1000 simulations were performed. In each simulation, the same dataset was analyzed with analysis of variance (ANOVA), analysis of covariance (ANCOVA) with p covariates, and ANCOVA with the PCs, which replaced the covariates. The null hypotheses (H_0) and alternative hypotheses (H_A) of the study were

$$H_0;\ \mu_1 = \mu_2 = ... = \mu_t$$

and $$H_A;\ \mu_i \neq \mu_j \text{ at least for one pair,}$$

where, μ_i are adjusted treatment means. The conditional statistical power was calculated as the percentage of rejections of the H_0 in the 1000 simulations within each experimental setup, at $\alpha = 0.05$. Values obtained for the power of the tests of (i) ANOVA, (ii) ANCOVA with p covariates, and (iii) ANCOVA with the PCs as covariates were compared to assess the impact of using PCs with (i) increase of correlation between covariates, (ii) increase of the number of covariates, (iii) increase of treatment mean difference, (iv) increase of the number of replicates, and (v) increase of the number of treatments.

3 Results and Discussion

This simulation study examined the possibility of incorporation of covariates through PCs in the analysis of covariance. In this case, a PC acts as a composite covariate. The composite covariate limits the loss of DF while explaining as much as possible of the variance explained by the individual covariates. The associated gain is particularly significant when the sample size is small and the number of covariates is large. As per the results, with the increase of correlation among covariates, treatment mean difference and number of covariates, the percentage of rejection of H_0 in 1000 simulations (power of the test) increases in ANCOVA with PCs, compared to ANOVA and ANCOVA with p covariates, when there are fewer number of observations.

Values of power of the test for four methods in analysis of covariance: (i) without any covariate (ANOVA), (ii) with p covariates, (iii) with one PC, and (iv) with two PCs; under three treatments and three replicates are shown graphically in Figure 1. It demonstrates that, with the increase of correlation among covariates, power of the test increases in ANCOVA with one PC and ANCOVA with two PCs, compared to ANOVA and ANCOVA with p covariates. It is also clear that, with the increase of treatment mean difference, power of the test increases in ANCOVA with one PC and ANCOVA with two PCs, compared to ANOVA and ANCOVA with p covariates, when there are three treatments and three replicates. The same pattern was observed with the increasing number of covariates too.

Figure 2 shows the power of the test of the four methods of analysis: ANOVA, ANCOVA with p covariates, and ANCOVA with one PC and ANCOVA with two PCs conditional on number of treatments and number of replicates and level of correlation among covariates, when the treatment mean difference is five and the number of covariates is four. As per the observations, power of the test of the four methods of analysis is higher when the number of treatments is five, compared to three treatments (Figure 2: Graphs A and B, C and D, E and F). Further, power of the test in ANCOVA with p covariates is higher, compared to ANCOVA with one PC and ANCOVA with two PCs, when the number of treatments is five. However, the power of the test in AN-

COVA with p covariates is lower, compared to ANCOVA with one PC and ANCOVA with two PCs, when the number of treatments is three (Figure 2: Graphs A, C, E).

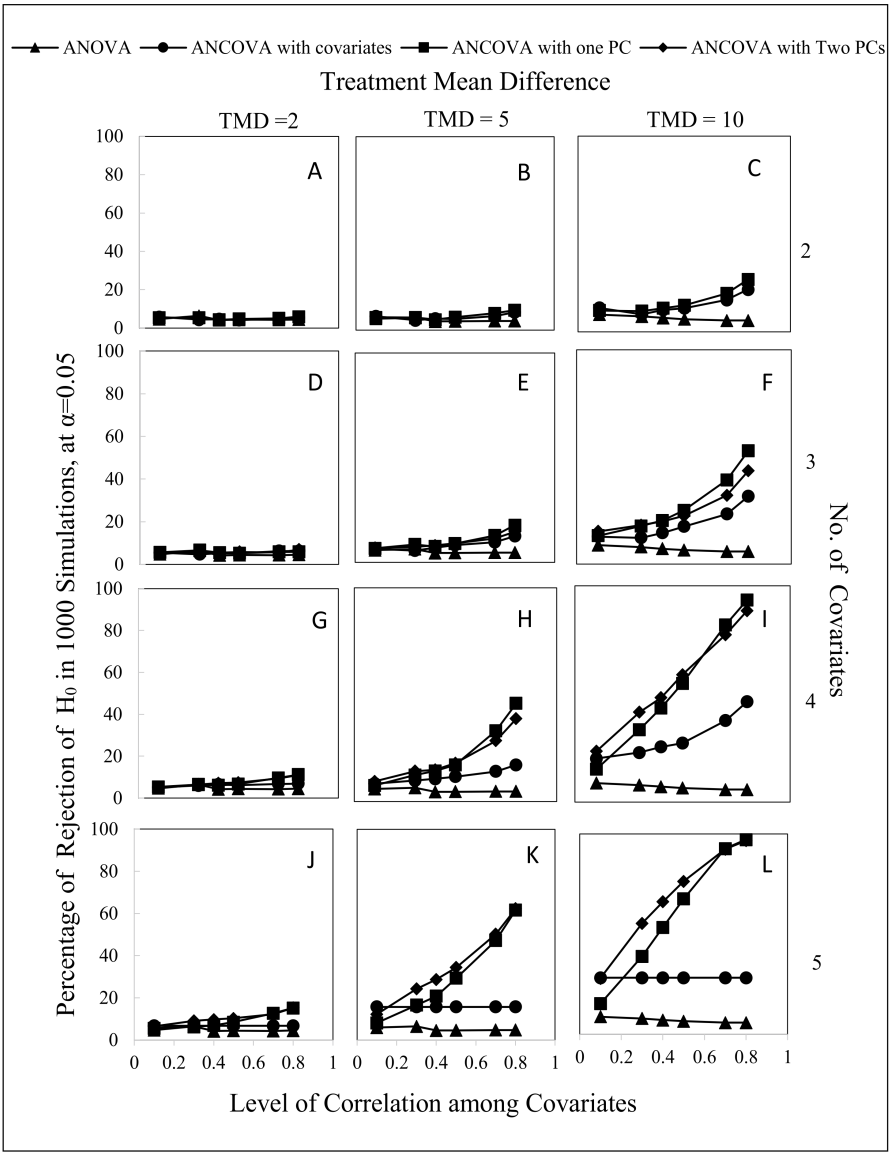

Fig. 1: Percentage of rejection of H_0 in 1000 simulations (power of the test) at α=0.05 with four methods in analysis of covariance, (i) without any covariate (ANOVA), (ii) with p covariates, (iii) with one PC, and (iv) with two PCs; against levels of treatment mean difference, no. of covariates, and level of correlation among covariates under fixed three treatments and three replicates. (TMD: Treatment Mean Difference).

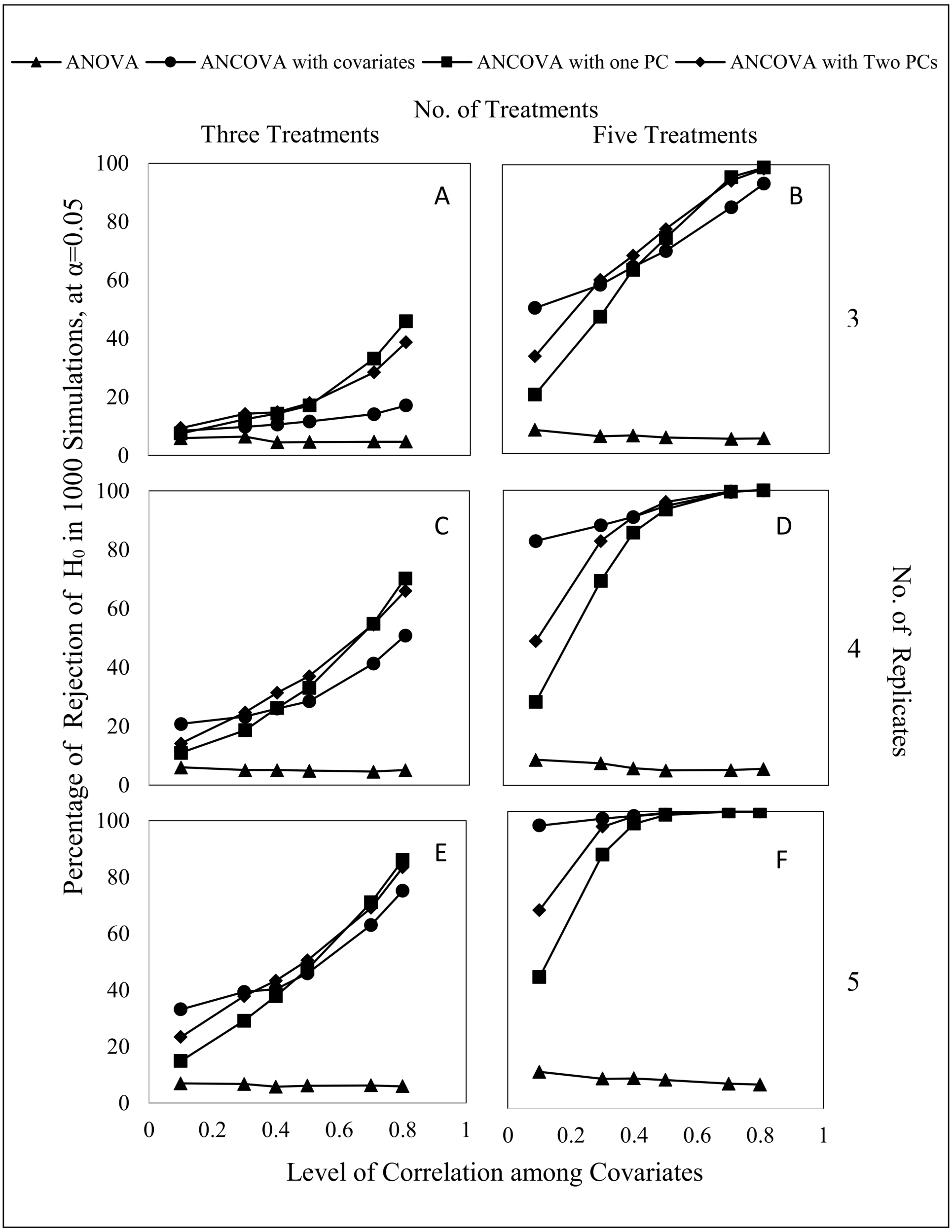

Fig. 2: Percentage of rejection of H_0 in 1000 simulations (power of the test) at α=0.05 with four methods in analysis of covariance, (i) without any covariate (ANOVA), (ii) with p covariates, (iii) with one PC, and (iv) with two PCs against differing levels of no. of treatments, no. of covariates and level of correlation among covariates under fixed four covariates and the treatment mean difference of five.

Figure 3 shows the power of the test of four methods of analysis: ANOVA, ANCOVA with p covariates, and ANCOVA with one PC and ANCOVA with two PCs, conditional on number of replicates and number of covariates when the number of treatments is three and the treatment mean difference is five. As per the results, power of the test is increasing with the increase of the level of correlation among covariates. The power of the test of ANCOVA with p covariates is increasing with the increase of number replicates as well as with the increase of number of covariates. Compared to five replicates, when the number of replicates is three, power of the test in ANCOVA with one PC and ANCOVA with two PCs is significantly higher in higher correlation levels than power of the test in ANCOVA with p covariates (Figure 3: Graphs J and L).

The power of the test of four methods of analysis: ANOVA, ANCOVA with p covariates, ANCOVA with one PC and ANCOVA with two PCs, with three treatments, when there are five covariates and correlated to form two PCs is expressed in Figure 4. A significant increase of power of the test was observed in ANCOVA with two PCs compared to the ANCOVA with one PC, when the level of correlation among correlated covariates is increasing. At higher correlation levels, power of the test in ANCOVA with two PCs is higher compared to power of the test in ANCOVA with p covariates and ANCOVA with one PC, when the number of observations is limited (Figure 4: Graphs A, B and C). With the increase of number of replicates, power of the test in ANCOVA with p covariates and ANCOVA with two PCs is increased while there was no remarkable increase in the power of the test in ANCOVA with one PC (Figure 4: Graphs B, E and H).

The results have revealed that the use of principal components can lead to a significant gain in the power of the test in ANCOVA. It is obvious, the gain depends on the variance explained by the principal component. With the increase in the level of correlation among covariates, the variance explained by the principal component increases. Accordingly, power of the test in ANCOVA with one PC and ANCOVA with two PCs increases with the increase of the level of correlation among covariates. With the increase of treatment mean difference, the average variation between groups becomes large compared to the average variation within groups, and thus the likelihood of rejection of H0 increases. Therefore, in all the situations, the number of rejections of H0 in 1000 simulations increases with the increase of treatment mean difference.

The addition of each covariate increases the power of the test in ANCOVA with p covariates, as they could correct for potential bias coming from baseline covariates. However, adding covariates in the analysis comes with a cost in DF. When there is a substantial number of observations, the loss of few DF for the covariates is not a serious issue. Even with the loss of DF, the addition of more covariates improves the precision of the analysis compared to the use of principal components, as the variance explained by all covariates might not be explained by the principal components, even with the highly correlated covariates.

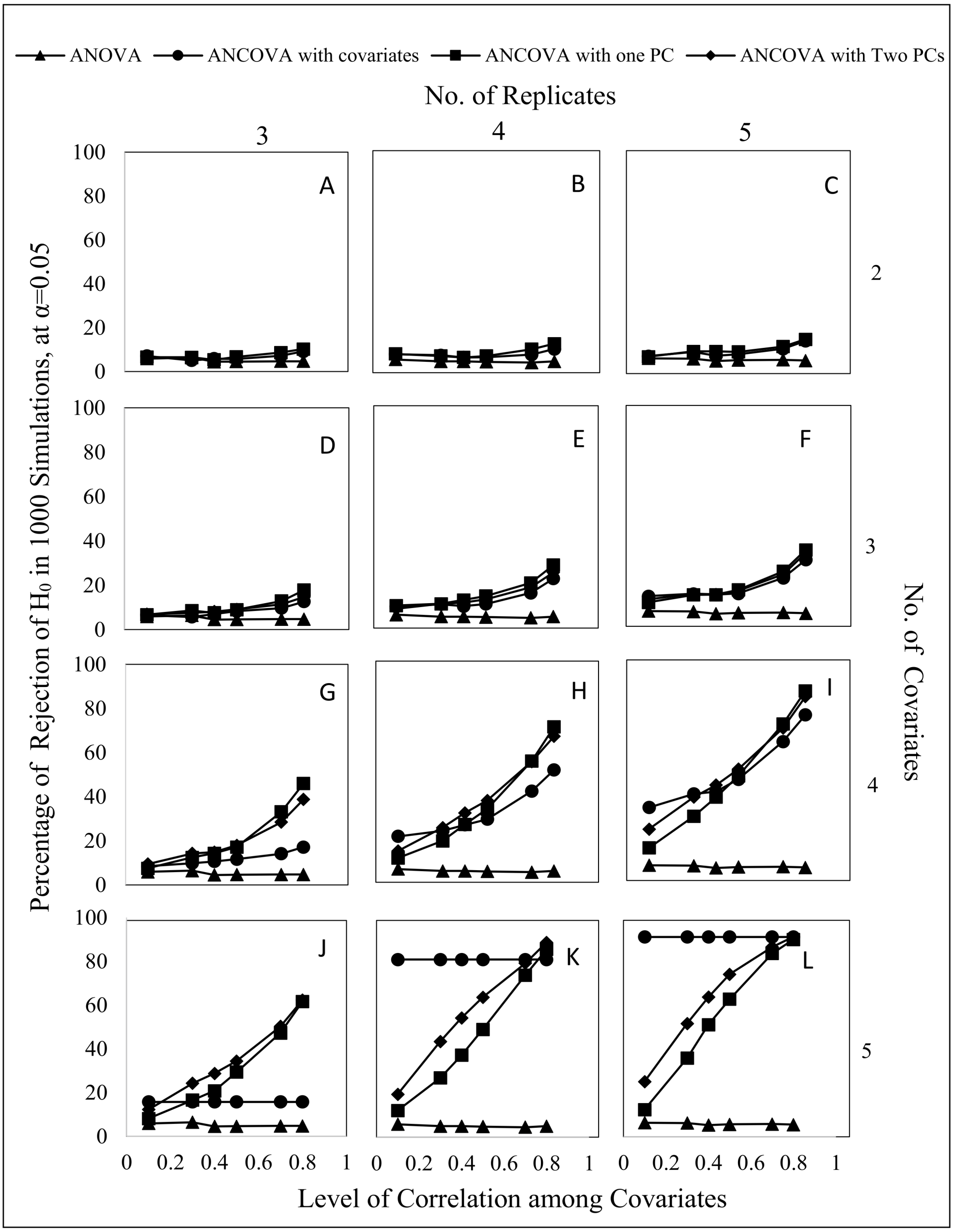

Fig. 3: Percentage of rejection of H_0 in 1000 simulations (power of the test) at α=0.05 with four approaches in analysis of covariance: without any covariate (ANOVA), with p covariates, with one PC and with two PCs with three treatments when the treatment mean difference is five. Estimates are given at differing levels of no. of replicates, no. of covariates and level of correlation among covariates

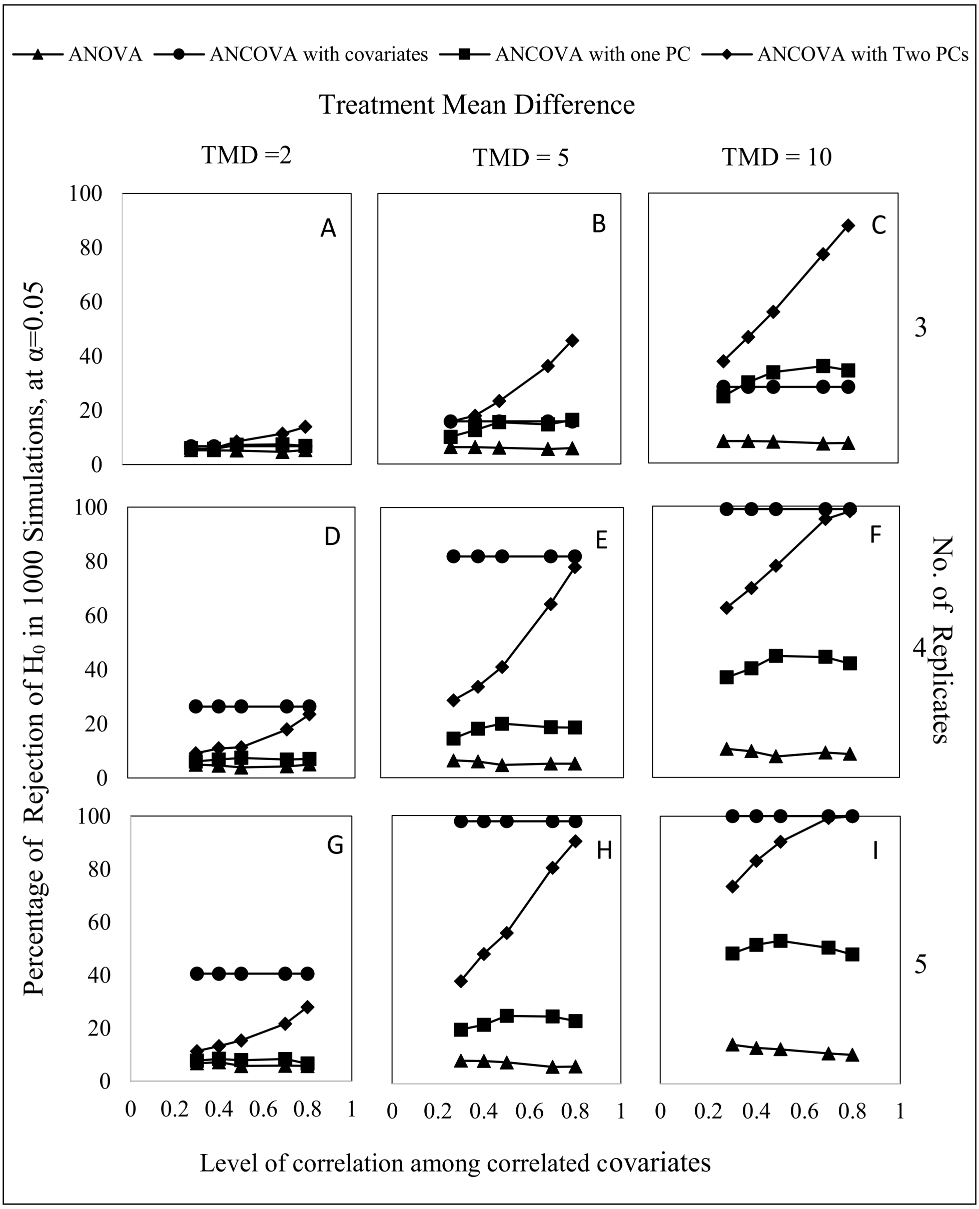

Fig. 4: Percentage of rejection of H0 in 1000 simulations (power of the test) at α=0.05 with four approaches in analysis of covariance: without any covariate (ANOVA), with p covariates, with one PC and with two PCs with three treatments when the five covariates are correlated to form two PCs. Estimates are given at differing levels of no. of replicates, treatment mean difference and level of correlation among correlated covariates

However, when the number of observations is limited, there can be a substantial reduction of the precision even with a loss of one DF. When the covariates are replaced with the PCs, a significant increase in the power of the test was observed when there is a limited number of observations, as it can preserve the DF. Therefore, when there is a limited number of observations, it is more appropriate to use a principal component as a composite covariate, even though the variance explained by the principal component is less, compared to all covariates, as preserving the DF might increase the precision of the analysis. Yet, the decision on the number of PCs to be included in the analysis should be made based on the correlation pattern among covariates. When the covariates are correlated to form 2 PCs and if only one PC is considered in the analysis, it would produce less precise results, especially when there is a limited number of observations.

The use of principal components as the composite covariates significantly improve the power of ANCOVA when the constructed PCs provide the maximum information that all the covariates could provide. When there are a number of inter-correlated covariates, use of PCs in ANCOVA could have a greater power. However, when the correlation among covariates is very low, use of a PC as a composite covariate can reduce the power of the test, especially when there is a substantial number of observations. In such cases, only the most contributing covariates could be incorporated as individual covariates. In such cases, the most contributing covariates could be incorporated as individual covariates. In case some covariates are correlated and others are not, then those related ones could be replaced by PCs while non-related ones could be used as individual covariates. However, studies should be done to verify the expected results.

4 Conclusion

The use of PCs as composite covariates instead of all covariates significantly increases the power of the test, when there is a limited number of observations, and the level of correlation among covariates is high. Therefore, it can be suggested to use the PC approach in incorporating covariates in the analysis of covariance if there is a number of inter-correlated covariates, especially when the number of observations is limited. It is recommended to investigate the performance of this PCA approach with real data to verify the findings.

References

Branders S., Pereira A., Bernard G., Ernst M., Dananberg J. and Albert A. (2021) Leveraging historical data to optimize the number of covariates and their explained variance in the analysis of randomized clinical trials, *Statistical Methods in Medical Research*, 31(2):240-252. DOI:10.1177/09622802211065246

Clements L., Kimber A. C. and Biedermann S. (2022) Multiple Imputation of Composite Covariates in Survival Studies, *Stats*, 5(2):358-370. DOI: 10.3390/stats5020020

Crager M. R. (1987) Analysis of covariance in parallel-group clinical trials with pretreatment baselines, *Biometrics*, 43(4): 895-901. PMID:3427174.DOI: 10.2307/2531543

Dharmawardena J. S. N. P., Thattil R. O. and Samita S. (2016) Adjusting variables in constructing composite indices by using principal component analysis: illustrated by Colombo district data, *Tropical Agricultural Research*, 27(1): 95–102. DOI: 10.4038/tar.v27i1.8157

Huitema B. (2011) *The Analysis of Covariance and Alternatives: Statistical Methods for Experiments, Quasi-Experiments, and Single-Case Studies, Second Edition.*, John Wiley and Sons, Hoboken, New Jersey, United States

Jolliffe, I. T. (2002) *Principal component analysis, 2nd edition. Springer Series in Statistics.*, New York: Springer-Verlag, New York.

Li T., Zhang H., Yuan C., Liu Z. and Fan C. (2012) A PCA-based method for construction of composite sustainability indicators. *The International Journal of Life Cycle Assessment*, 17(5): 593-603. DOI: 10.1007/s11367-012-0394-y

Mefford J. and Witte J. S. (2012) The Covariate's Dilemma, *PLoS Genetics*, 8(11): e1003096. DOI: 10.1371/journal.pgen.1003096

Raab G. M., Day S. and Sales J. (2000) How to select covariates to include in the analysis of a clinical trial, *Controlled Clinical Trials*, 21(4):330-342. DOI: 10.1016/S0197-2456(00)00061-1

Senna L. D., Maia A. G. and Medeiros J. D. F. (2019) The use of principal component analysis for the construction of the Water Poverty Index, *Brazilian Journal of Water Resources*, 24: 1–14. DOI: 10.1590/2318-0331.241920180084

Shieh G. (2020) Power Analysis and Sample Size Planning in ANCOVA Designs, *JPsychometrika*, 85(1):101-120. DOI: 10.1007/s11336-019-09692-3

Song M. K., Lin F. C., Ward S. E. and Fine J. P. (2013) Composite variables: when and how, *Nursing Research*, 462(1):45-9.DOI:10.1097/NNR.0b013e3182741948

Yang R. and Juskiw P. (2011) Analysis of covariance in agronomy and crop research, *Canadian Journal of Plant Science*, 91(4): 621-641. DOI: 10.4141/cjps2010-032