Nonparametric Test of Interaction Effect for 2^2-Factorial Design with Unequal Replicates: Case of Poisson-Normal Multivariate Data

Mara Sherlin D. Talento
Institute of Statistics, UP Los Banos

Marcus Jude P. San Pedro
De La Salle University

Erniel B. Barrios
School of Statistics, UP Diliman

Multivariate Analysis of Variance (MANOVA) is fairly robust to the normality and constant variance assumptions provided that the data is generated from a balanced design. Issues with hypothesis-testing arises when error-distribution is non-normal or when the data is generated from an unbalanced design. We propose a nonparametric method of testing interaction effect of the two-factor factorial design with multivariate response and possibly highly unbalanced replicates. Simulation studies indicated that the test is correctly-sized and increasing power with increasing effect-size, and increasing sample size. The parametric test based on MANOVA is incorrectly sized with unbalanced design and error distribution is not normal.

Keywords: two-factor factorial design, unbalanced replicates, nonparametric, multivariate data, Poisson and Normal data

1. Introduction

One of common treatment designs in experimental studies is the factorial design. Its most basic case is the two-factor factorial. Difficulty arises when the study involves more than one parameter. In particular, experiments conducted by the different fields in biological science include as much variables to come up with a solution or understanding on a certain phenomenon. One example is the biological composition of bacteria or cell exposed to certain types of growth factor. In order to test the significance of such factors for more than one variable of interest, Multivariate Analysis of Variance (MANOVA) is known to be a powerful tool.

As in the case of the univariate Analysis of Variance (ANOVA), certain assumptions, such as normality are to be satisfied in MANOVA. However, we know that ANOVA is fairly robust to the normality assumption and homogeneity of variances among experimental units, provided that the design is balanced, i.e., we have equal sample sizes. These types of assumptions are also necessary for the MANOVA setting. However, there are cases where experiments have failed measurement or incorrect experimental set-up which results to unequal replication. When these phenomena occurred (i.e., unbalanced design), test for normality is required. This paper looked at the distortion of the variance-covariance structure, i.e., inducing a count realization and the unbalanced scenario, which is inevitable in most of uncontrolled research set-up.

This heteroscedasticity is one of the major barriers in performing two-way MANOVA. Xu (2015) showed theoretically and computationally that parametric bootstrap outperformed Lawley-Hotelling trace and approximate Hotelling T^2 tests when the design is not balanced (i.e., unequal sizes and unequal cell covariance matrices). However, the test assumed symmetric distribution of errors.

Konietschke et al. (2015), on the other hand, compared the nonparametric bootstrap, parametric bootstrap, Walds' and Wilk's tests in testing interactions for general MANOVA. Their simulations include the setting of both symmetric and skewed distributions (i.e., Chi-squared distribution), as well as for balanced and highly unbalanced sample size allocations. Their results showed that parametric bootstrap had the best performance in terms of accuracy even for some skewed distributions. However, for cases with relatively small sample sizes (i.e., highly unbalanced) and negative pairing none of their tests provided good result. One solution that they are looking into is the permutation procedure.

In line with that, Friedrich et al. (2017) noted that in general, the Wald-type statistic (WTS) performs asymptotically well for the case of unequal covariance matrices and non-normal multivariate observations. The recent paper discussed correcting type-I error of small samples by the means of a permutation procedure through simulation studies. However, the unbalanced case for non-normal multivariate observations also violates the orthogonality of the MANOVA, which leads to inappropriate asymptotic assumptions of the WTS.

This paper was motivated by an experiment that involved determining the significance of packaging and preservative to biochemical properties of pineapple chunks, such as oxygen, carbon dioxide and micro-bacterial counts. Before proceeding to the significance of two factors (packaging and preservative), it is important to determine if interaction effect is present among the biochemical properties. Hence, the focus of this paper is to propose a nonparametric approach for a skewed multivariate set up, specifically the Poisson-Normal multivariate set-up with unbalanced case.

To provide a better understanding of the testing procedure, we note further that most of the established testing procedure assumes large sample approximation of the distribution. However, samples gathered from an experimental set-up are controlled. Thus, we propose a replication procedure of the data to provide an evidence against the null hypothesis.

2. Model and Hypotheses

This model adapted the data generating process of Xu (2015) for two-factor factorial treatment design laid out in completely randomized design but assuming Poisson-Normal distribution. Hence for two factors A and B with 2 levels each, i.e., i,j =1,2, and let $n = [n_{11}, n_{12}, n_{21}, n_{22}]$ and $\sum_{ij}$ be the covariance matrices

$$y_{ijk} = \mu_{ij} + \Sigma_{ij}^{1/2} e_{ijk};\ k = 1,2,\ldots,n_{ij} \text{ and } \mu_{ij} = \alpha_i + \beta_i + \frac{\alpha\beta_{ij}}{\gamma}.$$

The y_{ijk} in this set-up is the multivariate realization from three different random variables (i.e., p=3), with one response having a count response and two normal random variates. The $\boldsymbol{\alpha}_i$ and the $\boldsymbol{\beta}_i$ are the two treatment effects with two levels where $\boldsymbol{\alpha}_2 = \omega_1 * \boldsymbol{\alpha}_1$ and $\boldsymbol{\beta}_2 = \omega_2 * \boldsymbol{\beta}_1$. The ε_{ijni} follows that of a multivariate Normal Distribution but induced with a Poisson Distribution generated from the 'PoisNor' R package. The interaction term is a function of γ which has masking effect on the presence of interaction. Hence, higher value for γ means lower interaction effect.

The model above specifically answers hypothesis for two-factor factorial. This paper is only concerned with the interaction of two factors. That is, H_0 (AB): $\{(P_2 \otimes P_2 \otimes I_3) = 0\}$, where P_2 is the 2-dimensional centering matrix for the two factors (A and B) or simply $H_0(AB)$: $\boldsymbol{\alpha\beta}_{ij} = 0$.

3. The Test Statistic and Nonparametric Bootstrap

Test Statistics Based on Eigenvalues

1. Wilks (1932) – useful whenever the distributional assumptions are met. The formulation is $\Lambda_W = \prod_{j=1}^{d} (1+\phi_j)^{-1}$.
2. Pillai-Bartlett trace (Pillai 1955) – Is a robust statistic for smaller sample sizes, unequal number of cells (treatment combinations), and unequal variances. The formulation is $\Lambda_{PB} = \sum_{j=1}^{d} \frac{\phi_j}{1+\phi_j}$.

3. The Lawley-Hotelling trace is another statistic which is generally similar to Wilks Λ_W.

Nonparametric bootstrap approach

Using nonparametric bootstrap (Lancaster, 2003), randomization of vectors will be selected independently per cell with replacement where each bootstrap resample is a complete set of balanced design using the minimum sample size as the size allocation, that is, $B^* = [y_{11}, y_{12}, y_{21}, y_{22}]$.

This paper utilized the properties of Wilks statistic and observe the properties whenever distributional assumptions are not met. The bootstrap confidence interval (CI) will be constructed and whenever it does not contain $\chi^2_{\alpha,p}$ which in this case 7.81 (p=3, number of variables), the null hypothesis is rejected. Chi-square is used as criterion since Wilk's Λ_W follows $\chi^2_{\alpha,p}$ and is always positive.

4. Simulation Setting

In order to evaluate the test for interaction effect, first treatment effect (α_1) will be fixed to 5 and effect of second treatment (β_1) was allowed to vary. The first treatment has higher effect than the second treatment thus percentage contribution of the effect of $\boldsymbol{\alpha}$ was fixed to have higher percentage (i.e., negative pairing). Konietschke et al. (2015) showed that bootstrap has low accuracy among cases with negative pairings, hence, this paper also assessed the level of negative pairings that bootstrap can tolerate. The masking parameter of interaction effect (γ) will also be examined as well as increasing the effect of the treatment effects. The sample size allocation has 4 settings for (1) equal sizes with relatively high values [9,9,9,9], (2) unequal sizes with relatively high values [9,9,8,7], (3) unequal sizes with relatively small values [5,8,5,7] and (4) highly unbalanced sizes [9,9,8,3]. The simulation settings are summarized in Table 1.

Table 1. Settings of the Data for Power Analysis

β_i	Effect of α_i and β_i (ω_1; ω_2)	Treatment interaction Coefficients $\left(\frac{1}{\gamma}\right)$	Sample sizes (n)
5,7,12	10%; 5% (slightly negative pairs) 40%; 30% (highly negative pairs)	0.10, 0.25, 0.50	[9,9,9,9] [9,9,8,7] [5,8,5,7] [9,9,8,3]

The simulation has 500 bootstrap resamples in 250 replicates to determine the power and size. The probability of making a correct decision is the power of the test, that is P (*Rej* H_0 | H_0 *is false*). And the size of the test is the probability of committing a Type-I error is denoted as the probability of rejecting a true null hypothesis, that is, P (*Rej* H_0 | H_0 *is true*). In evaluating the power of the test, we count the number of bootstrap CI not containing $\chi^2_{\alpha=0.05,p=3}$ whenever there is an interaction effect out of the 250 replicates while the evaluation of the size for the proposed nonparametric test is whenever the bootstrap confidence interval (CI) does contain the $\chi^2_{\alpha=0.05,p=3}$.

5. Results and Discussion

Table 2 shows the size of the test or the probability of rejecting true null hypothesis. For the proposed test, the probability of rejecting null hypothesis, when in fact there is no interaction, is 0 for equal sample sizes. The size of the proposed test is also 0 for highly unbalanced sizes, even with 3 replicates. While the test is correctly-sized, however, its convergence is only 99.94%. Generally, the proposed test is within the acceptable allowable error which is 5%. On the other hand, the adjusted MANOVA has inflated size of the test for case of slightly negative pairing or (i.e., pertaining to effects of $\boldsymbol{\alpha}_i$ and $\boldsymbol{\beta}_i$). This might be due to the Poisson random variable which induced heteroskedasticity. The case of highly negative pairing or $\omega_1 = 40\%$ and $\omega_2 = 25\%$ has more conservative output. In fact, for $\boldsymbol{\beta} < 12$, the case of equal sample sizes and case of unequal sizes with relatively small values have acceptable size of the test. However, majority of the cases using adjusted MANOVA yielded inflated size of the test.

Table 3 shows the power of the test or the probability of rejecting false null hypothesis when ω_1 is 10% and ω_2 is 5% (i.e., case of slight negative pairing). For both proposed test and adjusted MANOVA, the case with treatment interaction coefficient equal to 0.5 has ideal power (see Table 3). For the proposed test, it can no longer detect false null hypothesis when treatment interaction coefficient equal to 0.1 and $\boldsymbol{\beta} = 5$ (i.e., low treatment effect) but for $\boldsymbol{\beta} = 12$, all cases have relatively high power except for highly unbalanced case with 3 as the minimum sample size. Generally, adjusted MANOVA has higher power over the proposed test but based on Table 2, adjusted MANOVA also has higher probability of rejecting true null hypothesis.

Table 4 shows the power of the test or the probability of rejecting false null hypothesis when ω_1 is 40% and ω_2 is 25% (i.e., case of highly negative pairing). As we compare Table 3 with Table 2, highly negative pairs have more conservative output than that of slightly negative pairing.

Based on Table 4, the proposed test has very low power in detecting the interaction effect when coefficient is set to 0.1 and when ω_1 is 40% and ω_2 is 25%. However, the power increases as treatment effect also increases.

Generally, the case of highly unbalanced sample sizes with 3 as the minimum sample size (i.e., setting 4) has the lowest power among all settings and the case of balanced design (i.e., equal sample sizes) has the highest power. Furthermore, power increases as coefficient of interaction effect also increases.

Table 2. Size of the Proposed Test and Adjusted MANOVA for Testing the Interaction Effect of Two Factors with Poisson-Normal Multivariate Data.

β	***sample size allocation***	10%-5%		40%-25%		***convergence***
		proposed test	adjusted MANOVA	proposed test	adjusted MANOVA	
5	setting 1 [9,9,9,9]	0	0.056*	0	0.036	*100%*
	setting 2 [9,9,8,7]	0.004	0.06*	0.004	0.056*	*100%*
	setting 3 [5,8,5,7]	0	0.06*	0.004	0.044	*100%*
	setting 4 [9,9,8,3]	0	0.064*	0	0.064*	*99.94%*
7	setting 1 [9,9,9,9]	0	0.054*	0	0.044	*100%*
	setting 2 [9,9,8,7]	0.004	0.06*	0.004	0.060*	*100%*
	setting 3 [5,8,5,7]	0.005	0.06*	0.004	0.036	*100%*
	setting 4 [9,9,8,3]	0	0.072*	0	0.072*	*99.94%*
12	setting 1 [9,9,9,9]	0	0.048	0	0.052*	*100%*
	setting 2 [9,9,8,7]	0.004	0.056*	0.004	0.052*	*100%*
	setting 3 [5,8,5,7]	0.005	0.06*	0.004	0.036	*100%*
	setting 4 [9,9,8,3]	0	0.06*	0	0.064*	*99.94%*

**size of the test more than pre-set allowable error of 5%*

Table 3. Power of the proposed test and adjusted MANOVA for testing the interaction effect of two factors with Poisson-Normal multivariate data when is ω_1 10% and ω_2 is 5%.

β	sample size	coefficient					
		0.1		0.25		0.5	
		proposed test	adjusted MANOVA	proposed test	adjusted MANOVA	proposed test	adjusted MANOVA
5	[9,9,9,9]	0.24	0.744	1	1	1	1
	[9,9,8,7]	0.116	0.536	1	1	1	1
	[5,8,5,7]	0.116	0.404	0.885	0.992	1	1
	[9,9,8,3]	0	0.156	0.324	0.792	1	1
7	[9,9,9,9]	0.692	0.948	1	1	1	1
	[9,9,8,7]	0.472	0.864	1	1	1	1
	[5,8,5,7]	0.336	0.668	0.995	1	1	1
	[9,9,8,3]	0.008	0.296	0.868	0.972	1	1
12	[9,9,9,9]	1	1	1	1	1	1
	[9,9,8,7]	0.988	1	1	1	1	1
	[5,8,5,7]	0.896	0.992	1	1	1	1
	[9,9,8,3]	0.24	0.756	1	1	1	1

Table 4. Power of the Proposed Test and Adjusted MANOVA for testing the Interaction Effect of Two Factors with Poisson-Normal Multivariate Data when ω_1 is 40% and ω_2 is 25%

β	sample size	coefficient					
		0.1		0.25		0.5	
		proposed test	adjusted MANOVA	proposed test	adjusted MANOVA	proposed test	adjusted MANOVA
5	[9,9,9,9]	0.028	0.228	0.564	0.912	1	1
	[9,9,8,7]	0.012	0.152	0.348	0.792	1	1
	[5,8,5,7]	0.012	0.132	0.12	0.564	0.93	0.996
	[9,9,8,3]	0	0.063	0.004	0.256	0.384	0.82
7	[9,9,9,9]	0.084	0.392	0.936	1	1	1
	[9,9,8,7]	0.032	0.332	0.832	0.988	1	1
	[5,8,5,7]	0.036	0.212	0.405	0.88	0.995	1
	[9,9,8,3]	0	0.084	0.056	0.496	0.904	0.988
12	[9,9,9,9]	0.488	0.888	1	1	1	1
	[9,9,8,7]	0.228	0.764	1	1	1	1
	[5,8,5,7]	0.216	0.548	0.975	1	1	1
	[9,9,8,3]	0	0.236	0.688	0.936	1	1

6. Application

Revenue of food industries depends on the shelf-life of their products, particularly those engaging in pineapple-manufacturing business. Some researchers are eyeing for the efficiency of 1-Methylcyclopropene (1-MCP) as preservative of some fruits, however, they want to determine if this preservative has some interaction with the type of packaging they are using (clamshell and cling wrap). Some biochemical properties like oxygen content and carbon content are some responses that might indicate the quality of pineapple. Researchers would also look at some growth factors like actino-bacteria (AB) - the higher the counts of these growth factors, the lesser the quality of the storage. These growth factors are known to be Poisson distributed, hence, normality and sphericity assumptions of usual MANOVA might not be achieved.

One pineapple-manufacturing company in the Philippines is interested in this area of study. However, during their experiments, some replicates spoiled and that resulted to the imbalance of their experimental design. Results are reported on the table below.

Table 5. Test of Normality and Sphericity

Response Variables	Replicates	Doornik-Hansen test	Mauchly's Sphericity test
Oxygen	9	0.0482*	<0.0001*
Carbon	9		
AB	3		

Based on Table 6, we have sufficient evidence to say that data did not come from multivariate normal distribution and the sphericity assumption was not satisfied. Hence, the proposed test and adjusted MANOVA was used in this study.

Table 6. Results of Interaction Effect test using the Proposed Test and the Adjusted MANOVA

Test of interaction effect	Confidence Interval	*p*-value
proposed test	[14.7387, 52.3758]	<0.0001*
adjusted MANOVA	n/a	0.0009746*

Using the proposed test, we have sufficient evidence to say that 1-MCP has an interaction effect with the type of packaging.

7. Conclusion and Recommendation

Theory on experimental design parametric test (F-test) requires equal replicates for treatment with factorial design for its robustness to normality assumption and orthogonality of the basis of vector space. Problem arises when the random variable is a count data and failure of measurement for some experimental unit, hence unequal replicates. Complexity of the problem upsurge when we have multivariate data with Poisson stochastic term. This setting will not only violate the normality assumption but also the sphericity of the variance-covariance matrix.

To test interaction effect on multivariate setting with unequal replicates, Wilk's lambda was estimated through bootstrap approach and empirical distribution of the statistic was obtained. The function for interaction effect is $\frac{\alpha\beta ij}{\gamma}$ where denominator masks the effect of interaction.

Graph of the data set showed parallel effect when random variable is normal and inherent slight interaction when it is Poisson even with no interaction present (Appendix 1). This might be due to nature of Poisson random variable (i.e. single parameter, λ), hence, difference on its mean would imply difference on the variance that would result to heterogeneity of all cells and imitating behavior of interaction effect.

Results have shown increasing power for increasing β, holding other factors constant. On the other hand, power also increases for increasing interaction effect holding other factors constant. Also, highly negative pairing of effects decreases the power of the test. Generally, all size of the proposed test for cases mentioned are lower than or equal to 0.5%.

Acknowledgement

We want to thank Ms. Meryl Bernardino for allowing access to the data used in the illustration..

References

FRANK K., BATHKE A., SOLOMON H.W. and MARKUS P., 2015, Parametric and nonparametric bootstrap methods for general MANOVA, *Journal of Multivariate Analysis*, 140, 291-301.

FRIEDRICH, S., BRUNNER, E., and PAULY, M., 2017, Permuting longitudinal data in spite of the dependencies, *Journal of Multivariate Analysis*, 153, 255-265. doi:10.1016/j.jmva.2016.10.004.

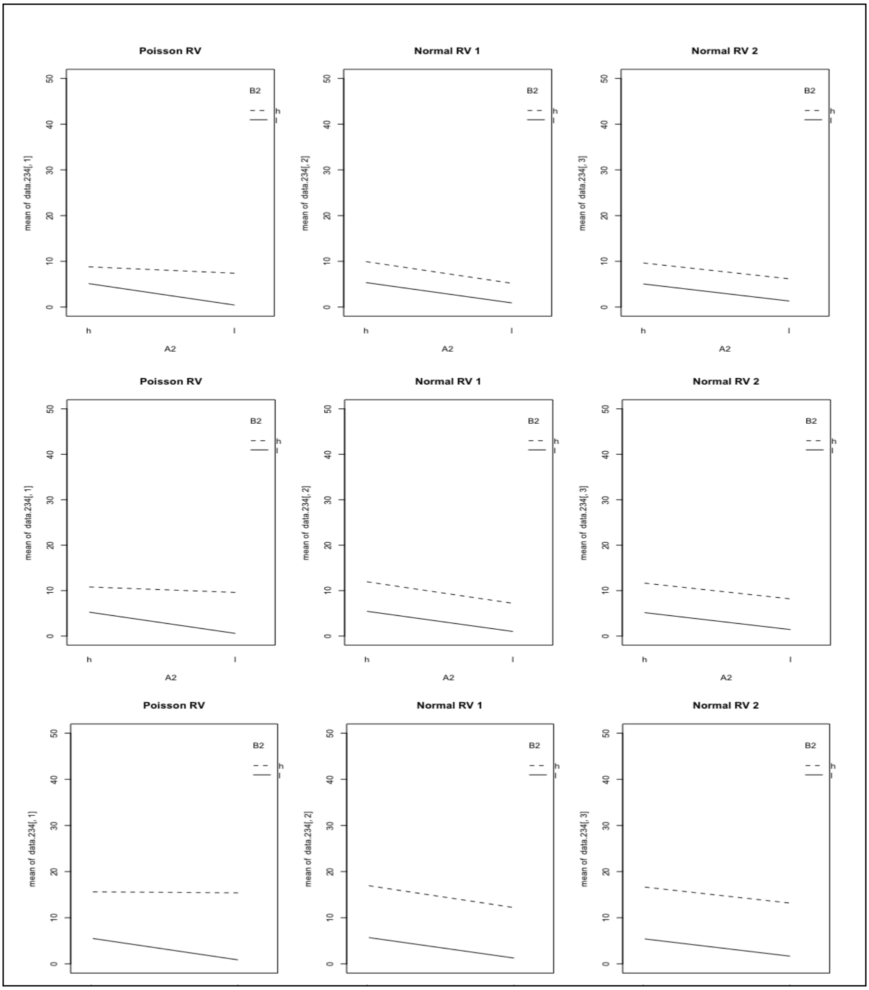
FRIEDRICH, S., KONIETSCHKE, F., and PAULY, M., 2017, GFD: An R package for the analysis of general factorial designs, *Journal of Statistical Software*,79. doi:10.18637/jss.v079.c01.

HARRAR, S.W. and BATHKE, A.C., 2008, Nonparametric methods for unbalanced multivariate data and many factor levels, *Journal of Multivariate Analysis*, 99, 1635-1664.

SEBER, G. A., 2015, *The Linear Model and Hypothesis: A General Unifying Theory*, Switzerland: Springer International Publishing.

XU, L.W., 2015, Parametric bootstrap approaches for two-way MANOVA with unequal cell sizes and unequal cell covariance matrices, *Journal of Multivariate Analysis*, 133, 291-303.

Appendix



Appendix 1. Interaction plot of 3 dependent variables for 10% - = 5% (top: , middle: ; bottom:) and sample sizes are [5,8,5,7]