STATISTICS IN TRANSITION new series, June 2022
Vol. 23 No. 2, pp. 163–183, DOI 10.2478/stattrans-2022-0022
Received – 01.08.2021; accepted – 06.04.2022

Advances on Permutation Multivariate Analysis of Variance for big data

Stefano Bonnini[1], **Getnet Melak Assegie**[2]

ABSTRACT

In many applications of the multivariate analyses of variance, the classic parametric solutions for testing hypotheses of equality in population means or multisample and multivariate location problems might not be suitable for various reasons. Multivariate multisample location problems lack a comparative study of the power behaviour of the most important combined permutation tests as the number of variables diverges. In particular, it is useful to know under which conditions each of the different tests is preferable in terms of power, how the power of each test increases when the number of variables under the alternative hypothesis diverges, and the power behaviour of each test as the function of the proportion of true alternative hypotheses. The purpose of this paper is to fill the gap in the literature about combined permutation tests, in particular for big data with a large number of variables. A Monte Carlo simulation study was carried out to investigate the power behaviour of the tests, and the application to a real case study was performed to show the utility of the method.

Key words: big data, MANOVA, permutation test, multivariate analysis.

1. Introduction

In many applications of the multivariate analyses of variance (MANOVA), the classic parametric solutions for testing hypotheses of equality in population means or multisample and multivariate location problems might not be suitable for various reasons. For instance, the strong and implausible assumptions of iid observations and multivariate normality are the main reasons for considering parametric methods neither flexible nor robust and consequently often unsuitable. Moreover, in the presence of big data with a high number of response variables, great attention should be paid when the number of response variables is larger than the sample sizes, because of the loss of degrees of freedom.

[1] Department of Economics and Management, University of Ferrara, Italy. E-mail: bnnsfn@unife.it. ORCID: https://orcid.org/0000-0002-7972-3046.

[2] University of Parma, Italy. E-mail: mlkgnt@unife.it. ORCID: https://orcid.org/0000-0001-7288-9636.

Even if there is not a unique definition, in statistics, a dataset is usually classified as “big data” if it represents a collection of informative data, extensive in terms of volume, velocity and variety, such that specific analytical technologies and methods are required for the extraction of value or knowledge (Baro et al., 2015). Big data are typical of many empirical disciplines such as biomedicine, economics, biology, ICT, education and research, financial services, social media, automotive industries, etc. (Özköse et al., 2015). Frequently, the high volume of big data depends on the multivariate nature of the dataset, due to the large number of variables. In addition, the variety of big data, due to the presence of different types of variables (quantitative and qualitative) and to the variability and heterogeneity of data, makes inferential problems more complex and requires robust and valid techniques to make inferences. For instance, in studies focused on social media, text, video, audio, and image data are jointly analysed. Hence, tests of hypotheses for big data must be addressed with appropriate methods that lead to reliable decisions, in short times and taking into account the variability and heterogeneity of the information.

A typical approach to variable oriented multivariate problems consists in the application of exploratory methods based on the dimensionality reduction such as principal component analysis (PCA) or factor analysis (FA) (Johnson and Wichern, 2007; Farcomeni and Greco, 2016). For two-sample multivariate testing problems, in the presence of numeric data, a typical solution is the Hotelling T-square test. These methods are based on strong assumptions such as the linearity of the relationships between variables or normality.

Linearity is a very strong and often unrealistic assumption. Normality is a reasonable assumption only with large sample sizes due to asymptotic properties of the statistics. Nevertheless, even in cases where linearity and normality are reasonable assumptions, especially in inferential problems, in the presence of many variables the estimation of a large number of unknown parameters, such as covariances or correlations, is required. Moreover, when the sample size is less than the number of variables, a problem related to the degrees of freedom arises and some typical parametric methods, such as the Hotelling T-square test, are not applicable.

In such problems, nonparametric methods are preferable because they do not require that the underlying probability law belongs to a given family of distributions and no parameters need to be estimated. In particular, permutation tests follow a distribution-free approach and are almost as powerful as parametric methods based on normality when this assumption is true but much more powerful when the true underlying distribution deviates from the Gaussian (Pesarin, 2001; Anderson, 2001).

Solutions for multivariate tests within the family of permutation methods consider the dependence between response variables without modelling it explicitly, and consequently without the need of estimating parameters or assuming linearity

(Pesarin and Salmaso, 2010a; Bonnini et al., 2014; Arboretti et al., 2018). Permutation solutions for multivariate location problems have been proposed and studied mainly in terms of power and robustness with respect to the underlying distribution, especially comparing their performance with that of the classic parametric tests (Pillar, 2013; Anderson, 2001; Pesarin, 2001). An interesting proposal is based on the combination of the univariate permutation tests of the marginal variables (Pesarin, 2001). Pesarin and Salmaso (2010a,b) proved that the power of the most commonly used combined permutation tests, with fixed sample size and divergent number of variables under the alternative hypothesis, tends to one in the two-sample problem.

According to the type of the combining function used, a different combined test is obtained. Hence a deep study with the goal of comparing different combined tests, especially for big data with a large number of variables, is important and suitable, in order to find the most powerful test under different scenarios. To the best of our knowledge, for the multivariate multisample location problem, a comparative study of the power behaviour of the most important combined permutation tests as the number of variables diverges is missing. In particular, it is useful to know under which conditions each of the different tests is preferable in terms of power, how the power of each test increases when the number of variables under the alternative hypothesis diverges and the power behaviour of each test as a function of the proportion of true alternative hypotheses.

The purpose of this paper is to fill this gap in the literature about combined permutation tests. The paper is organized as follows. Section 2 is dedicated to a review of the literature on the MANOVA problem. The method of combined permutation tests is described in Section 3. In Section 4 the results of a comparative simulation study are reported and discussed. In Section 5, the application of the method to a real case study is presented. Finally, the conclusions are in Section 6.

2. Literature review

The goal of several empirical studies is the comparison of two or more populations in the presence of multivariate response variables. Often, regardless of the number of factors, the problem consists in testing the significance of treatment effects or the presence of a shift in some location parameters. In what follows, the variation of population means is investigated using multivariate analysis of variance (MANOVA). To test whether there is a significant difference between group means, various parametric multivariate tests based on strong assumptions have been proposed. The most commonly used are the Hotelling T-square test (Hotelling, 1992), the test of Wilks (1932) and the proposal of Pillai (1955). The main assumptions of these tests are normality, constant variances and continuous responses. Moreover, these methods cannot be applied for big datasets when the number of response variables is greater than the sample size.

Nonparametric solutions have been proposed to overcome the limits of the tests mentioned above due to the lack of robustness with respect to the assumptions (Pesarin and Salmaso, 2010a; Bonnini et al., 2014; Pillar, 2013; Bonnini, 2016). For instance, Anderson (2001) introduced a nonparametric solution based on the permutation test for an ecological problem. The permutation test statistic was the Fisher F ratio obtained from a distance matrix, and the simulation results proved the appropriateness of the permutation test for both one-way and two-way MANOVA. Pillar studied the accuracy and power of permutation tests for MANOVA based on different test statistics. According to his study, the sum of squares between groups with the Euclidean distance was preferable to the Chord distance and the sum of Fs of univariate ANOVA. Moreover, the simulation study revealed that the permutation test was powerful also under heteroscedastic and with unbalanced samples.

In the literature, several works concerning applications of permutation tests for one-way and two-way MANOVA have been published. A non-exhaustive list includes the following papers: Mantel and Valand (1970), Mielke et al. (1976), Clarke (1993), Pillar and Orlóci (1996), Legendre and Anderson (1999), Mielke and Berry (1999), McArdle and Anderson (2001), Arboretti et al. (2018), Finch (2016). However, the extension of the permutation test for two-way MANOVA requires great attention in permuting the statistical units between groups. This is because the exchangeability condition is guaranteed only within the levels of one factor by considering the second factor as a block. Thus, constrained permutations are essential (Anderson, 2001). The two-sample multivariate problem has been frequently considered. See for instance Pesarin and Salmaso (2010), Polko-Zajac (2020), Bonnini and Melak Assegie (2019). Instead, the multi-sample case has been addressed by fewer authors (see Bonnini, 2016). In some cases permutation solutions for complex problems such as multiaspect tests (Polko-Zajac, 2019), directional alternatives (Bonnini et al., 2014; Arboretti and Bonnini, 2009), tests for categorical data (Arboretti and Bonnini, 2008; Bonnini, 2014) have been developed. In this paper, we focus on multi-sample location problems for numeric variables and nondirectional alternative hypotheses.

3. Methods

3.1. Multivariate permutation test

The permutation test is a distribution-free test based on the assumption of exchangeability under the null hypothesis (Pesarin, 2001). To apply the permutation principle, the sample data are partitioned into groups based on the treatment levels in an experimental study and pseudogroups in an observational study. To this end, the structure of the dataset for $S \geq 2$ independent samples and V-dimensional response is represented by:

$$\boldsymbol{Y} = \left\{Y_{igq} \middle| i = 1,2,\dots,n_g, g = 1,2,\dots,S, q = 1,2,\dots,V\right\} \tag{1}$$

The dataset $\boldsymbol{Y}$ takes values on the V-dimensional sample space Ω for which a σ-algebra $\mathcal{A}$ and a nonparametric family $\mathcal{P}$ of non-degenerate unknown distributions are defined, and supposed to be exchangeable.

Hypothesis testing based on the permutation approach requires a clear formulation of the null hypothesis. The null hypothesis in the MANOVA problem is defined as the equality of S multivariate (unknown) distributions:

$$H_o : \{P_1 = P_2 = \cdots = P_S\} = \{\boldsymbol{Y_1} \ =^d \boldsymbol{Y_2} \ =^d \dots =^d \boldsymbol{Y_S}\}. \tag{2}$$

Under homoscedasticity, the difference between the groups is due to a shift in location. Thus, the null hypothesis could be formulated as equality of group means for each response variable. Let $\boldsymbol{Y_g}$ be a V-variate numeric random variable such that $\boldsymbol{Y_g} = \boldsymbol{\mu} + \boldsymbol{\delta_g} + \boldsymbol{\varepsilon_g}$, with $\boldsymbol{\mu}$ vector of V unknown location parameters, $\boldsymbol{\delta_g}$, $g = 1,\dots,S$, vectors of V treatment effects and $\boldsymbol{\varepsilon_g}, g = 1,\dots,S$, exchangeable V-dimensional random vectors that follow an unknown probability distribution with equal variance-covariance matrix $\boldsymbol{\Sigma}$ and such that $E\left(\boldsymbol{\varepsilon_g}\right) = \boldsymbol{0}$.

The null hypothesis is:

$$H_o : \{\boldsymbol{\delta_1} \ = \boldsymbol{\delta_2} =, \dots, = \boldsymbol{\delta_S} = \boldsymbol{0}\} \tag{3}$$

A further decomposition of the null hypothesis with respect to the marginal distributions of the multivariate response can be considered. The multivariate hypothesis can be broken down into V partial null hypotheses:

$$H_o : \cap_{q=1}^{V}\left(\delta_{1q} =, .., = \delta_{Sq} = 0\right) \equiv \cap_{q=1}^{V} H_{oq} \tag{4}$$

where the intersection symbol means that the null hypothesis of the overall problem is true if all the V partial null hypotheses are true. Accordingly, with a similar approach, the alternative multivariate hypothesis H_1 of inequality in distribution may be represented as follows:

$$H_1 : \cup_{q=1}^{V} \overline{H}_{oq} \tag{5}$$

where the union symbol indicates that the alternative hypothesis is true if at least one partial null hypothesis is false and $\overline{H}_{oq}$ denote the negation of the q-th partial null hypothesis. It is worth noting that directional alternatives are also possible but the purpose of this paper is to focus on two-tailed multi-sample multivariate problems.

When the overall null hypothesis is true and the equality in distribution holds, the vector of V observations concerning a generic statistical unit comes from any of the S populations with equal probability. In other words, the exchangeability of units with respect to the populations/samples is satisfied. In order to determine the null distribution of the test statistic, all the possible assignments of the n units to the S samples can be considered. Without loss of generality, let us assume that the n_1 units of the first sample correspond to the first n_1 rows of the observed dataset $\boldsymbol{Y}$, the n_2 units of the second sample correspond to the next n_2 rows of the dataset, and so on, until the n_S units of the S-th sample that correspond to the last n_S rows of the dataset. Each possible assignment is equivalent to a permutation of the rows of the dataset or to resampling without replacement the n units with $n = n_1 + n_2 + \cdots + n_S$.

For computational convenience, instead of considering the exact test, based on all the $\frac{n!}{\prod_{g=1}^{S} n_g!}$ possible assignments of the n units to the S groups, a random sample of permutations is used according to the Conditional Monte Carlo method.

3.2. Partial tests

The application of the method of Combined Permutation Test to the permutation MANOVA presented above consists in carrying out one univariate permutation test for each partial hypothesis and in combining the p-values of the univariate tests. The dependence between the univariate partial test statistics, according to the permutation distribution, is taken into account in the resampling strategy by permuting the rows of the observed dataset instead of permuting the elements of each column independently of the other columns.

A suitable test statistic for each partial permutation test is the so-called Treatment Sum of Squares (SS_{Treat}), which depends on the deviations of the within-group sample means from the total sample mean. Hence, the q^{th} partial test statistic or equivalently the test statistic of the q^{th} partial test, with $q = 1,2, \dots, V$, is

$$T_q = \sum_{g=1}^{S} n_g \left(\bar{Y}_{gq} - \bar{Y}_{\cdot q}\right)^2 \tag{6}$$

with $\bar{Y}_{\cdot q} = \frac{\sum_g n_g \bar{Y}_{gq}}{\sum_g n_g} = \frac{\sum_g n_g \bar{Y}_{gq}}{n}$, where $\bar{Y}_{gq}$ represents the mean of the values of the q-th variable observed in the g-th sample.

The multivariate permutation distribution of the test statistic $\boldsymbol{T} = (T_1, T_2, \dots, T_V)$ under the null hypothesis is obtained through the following procedure:

1) compute the vector of observed values of $\boldsymbol{T}$ from the dataset $\boldsymbol{Y}$:

$$\boldsymbol{T_{obs}} = \boldsymbol{T}(\boldsymbol{Y}) = (T_{1,obs}, T_{2,obs}, \dots, T_{V,obs})$$

2) randomly permute the rows of the dataset (or reassign statistical units to groups) and compute the values of the test statistics as a function of the permuted dataset: $\boldsymbol{T}^p = \boldsymbol{T}(\boldsymbol{Y}^p)$
3) repeat step (2) R times independently and compute the permutation test statistics. Let $T_{q,r}^p$ be the value of the q-th partial test statistic related to the r-th permutation of the dataset $\boldsymbol{Y_r^p}$. Hence

$$\boldsymbol{T_r^p} = \boldsymbol{T}(\boldsymbol{Y_r^p}) = (T_{1,r}^p, T_{2,r}^p, \dots, T_{V,r}^p)$$

4) estimate the significance level function of the partial tests

$$\hat{\lambda}_{q,r}^p = \lambda(T_{q,r}^p) = \frac{\sum_{j=1}^R \mathrm{I}(T_{q,j}^p \geq T_{q,r}^p) + 0.5}{R+1} \tag{7}$$

with $r = 1, 2, \dots, R$, $q = 1, 2, \dots, V$, and $I(E)$ indicator function of E, which takes value 1 if E is true and 0 otherwise. The p-value of the q-th partial test is $\hat{\lambda}_{q,obs}^p = \lambda(T_{q,obs}^p)$.

3.3. Combination

According to the method based on the combination of dependent permutation tests, the test statistic for the overall problem is obtained by combining the p-values of the partial tests. The synthesis of the information provided by the partial tests regarding the marginal variables is provided by the application of a suitable combining function φ. Hence, the test statistic useful for the overall test, the multivariate analysis of variance, is

$$T_{comb} = \varphi(\lambda_1, \lambda_2, \dots, \lambda_V).$$

The proposal of combining p-values of partial tests in order to solve multivariate, multi-aspect, multi-strata tests, or other complex testing problems that can be broken down into partial univariate tests, appeared for the first time in the literature twenty years ago in Pesarin (2001) and was later studied and developed by several authors. For extended but not exhaustive reviews, see Pesarin and Salmaso (2010a) and

Bonnini et al. (2014). Since, for the combination of the partial tests, $\varphi(\cdot)$ must satisfy some simple, mild and easily attainable conditions, several different functions can be used and each of them corresponds to a different solution with specific properties within the family of combined permutation tests.

A suitable combining function $\varphi: (0,1)^V \to \mathbb{R}$ must satisfy the following properties:

1) $\forall(\lambda'_q, \lambda''_q)$ in (0,1), $\lambda'_q < \lambda''_q \Leftrightarrow \varphi(\dots, \lambda'_q, \dots) \geq \varphi(\dots, \lambda''_q, \dots)$ ceteris paribus (non-increasing monotony)
2) $\exists \lambda_q \epsilon \{\lambda_1, \lambda_2, \dots, \lambda_V\}$ s.t. $\lambda_q \to 0 \Leftrightarrow \varphi(\lambda_1, \lambda_2, \dots, \lambda_V) \to \bar{\varphi} < \infty$ (finite supremum)
3) $\forall \alpha \epsilon (0,1)$, $\exists T_{comb,\alpha} < \bar{\varphi}$ where $T_{comb,\alpha}$ is the test critical value (finite critical value)

The most popular combining functions in the literature of combined permutation tests are Fisher, Liptak and Tippett functions. The Fisher omnibus combining function is

$$T_F = -2 \sum_q log(\lambda_q) \tag{8}$$

where $log(x)$ denotes the natural logarythm of x. Liptak`s combining function is based on the transformation of the complement to one of the p-values through the inverse of the cumulative distribution function (or the quantile function) of the standard normal distribution:

$$T_L = \sum_q \Phi^{-1}(1 - \lambda_q) \tag{9}$$

where $\Phi(x) = P(X \leq x)$ with $X \sim \mathcal{N}(0,1)$. Tippett combination is based on an order statistic and considers, as observed value of the combined test statistic, the complement to one of the most significant p-value:

$$T_T = max_q\{1 - \lambda_q\} \tag{10}$$

Under the null distribution, if the V partial tests are independent and continuous, the Tippett function follows the uniform distribution in (0,1).

Without loss of generality, let us assume that the null hypotheses of the overall and partial problems are rejected for large values of the respective test statistics. It is trivial to show that all three combination rules defined above satisfy this condition. Given that the observed value of the combined test statistic is

$$T_{comb,obs} = \varphi(\hat{\lambda}^p_{1,obs}, \hat{\lambda}^p_{2,obs}, \dots, \hat{\lambda}^p_{V,obs}).$$

the p-value of the permutation MANOVA with the combined permutation test is given by

$$\hat{\lambda}_{comb,obs} = \lambda(T_{comb,obs}) \tag{11}$$

The three presented tests can have much different power behaviours under different conditions, hence a comparative analysis to deepen their properties, advantages and limits is important to support the analyst in the decision about which test to use based on the power.

4. Simulation study

The power behaviour of the three combined permutation tests defined in the previous section for the MANOVA problem was investigated through a Monte Carlo simulation study. Different scenarios, under the null and the alternative hypothesis, were considered in order to compare the power of the three proposals as a function of the sample sizes, of the number of samples, of the number of components of the multivariate response and of the proportion of true partial alternative hypotheses when H_0 is false.

Data were simulated according to the one-way MANOVA model. We considered multivariate datasets with two different sizes from the point of view of the number of responses: $V = 50$ and $V = 100$. With regard to the number of compared samples, $S = 3$ and $S = 5$ are the cases taken into account. Simulation study has been carried out generating data from V-variate normal random variables hence under the "probabilistic condition most favorable to the classic parametric tests" and under homoscedasticity. For all the S populations, the variance of each of the V components of the multivariate response and the correlation between any pair of variables was set equal to 1 and to 0.3 respectively. Hence, the $V \times V$ covariance matrix of each population is $\Sigma = [\sigma_{kq}]$ with $\sigma_{qq} = 1$, $q = 1,2,\dots,V$, and $\sigma_{kq} = 0.3$, $k \neq q \epsilon \{1,2,\dots,V\}$.

The number of simulated datasets and the number R of permutations were both equal to 1000. In the simulations, we considered the balanced design with size $n_1 = n_2 = \cdots = n_S = n$. The two sample sizes taken into account are $n = 10$ and $n = 30$. In the simulations, $\boldsymbol{\mu} = \mathbf{0}$. Let p be the proportion of true partial alternative hypothesis. Then, the V-variate normal distribution of the random variable that simulates data for the g-th sample $(g = 1,2,\dots,S)$ has a vector of means with $(1-p)\,V$ zeros and pV values equal to $\tau(g-1)$. Formally

$$\boldsymbol{\delta}_g = \tau(g-1)\binom{\mathbf{1}}{\mathbf{0}}$$

where $\mathbf{1}$ is a vector of pV elements equal to 1 and $\mathbf{0}$ is a vector of $(1-p)\,V$ elements equal to 0. To consider different shifts in the population locations, the simulations were carried out with $\tau = 0.5$ and $\tau = 1.0$. Moreover, the different proportions p of true alternative hypotheses used in the scenarios are $0.00, 0.05/0.06, 0.10, 0.20, 0.30, 0.40, 0.50, 0.70, 0.90, 1$. The first positive proportion in the list is 0.05 if $q = 100$ (5 true partial alternative hypotheses) and 0.06 if $q = 50$ (3 true partial alternative hypotheses). The significance level chosen in all the scenarios is $\alpha = 0.05$. All simulations were carried out with the R programming software version 4.1.0. Specific scripts were created by the authors for this purpose.

Table 1 shows the rejection rates of the tests under all different cases when the number of variables V is equal to 100. The performance of the tests under H_0 can be evaluated from the column corresponding to $p = 0.00$ (no true partial alternative hypotheses). It is evident that, in most cases, the rejection rates are either less than or very close to the nominal α level 0.05. The test based on the Tippet combination exceeds α more frequently than the others but the probability of wrong rejection of H_0 seems to be not far from 0.05, hence we can say that all the tests are well approximated.

When $p > 0$, the power behaviour of the tests can be assessed under H_1. Unbiasedness of all the tests is demonstrated because the rejection rates are greater under the alternative hypothesis than under the null hypothesis. Moreover, the greater the sample size the higher the power, thanks to the consistency of the tests. As expected, the power is increasing function of the shift of the population locations that depends on τ. Finally, the greater the number of samples the higher the rejection rates of the tests. Focusing on the effect of p on the estimated probability of rejecting H_0 when it is false, the increasing monotonic relationship is evident for all the tests. The growth rate of the power with respect to p is high and, when 100% of the partial alternative hypotheses is true, the rejection of H_0 is sure or almost sure.

From the comparative analysis, it emerges that the Liptak test is always the worst, except in the case in which all the partial alternative hypotheses are true. As said, in this scenarios, the power of all the combined tests tends to one and the tests are equivalent. In general the lower performance of the test based on the Liptak combination is evident and it is uniformly less powerful than the other permutation MANOVAs. This is consistent with Pesarin's (2001) statement about the preferability of other tests than Liptak, except for p=1. When the proportion of true partial alternative hypotheses is low, the combined test based on Tippett's rule is by far the best. Also this conclusion is not surprising, according to Pesarin (2001) but, in our simulation study, the extent of the difference in performance of the test based on Tippet's function can be evaluated. Moreover, according to this results, Tippet's combination is never less performant than the others, except in the first setting, when $S = 3$, $n = 10$ and $\tau = 0.5$ when $p \geq 0.90$, where the differences in the rejection rates of the various tests are negligible.

Table 1. Rejection rates of combined permutation tests for $V = 100$ and $\alpha = 0.05$.

S	**n**	**τ**	**φ**	**Proportion of true partial alternative hypotheses (p)**									
				0.00	**0.05**	**0.10**	**0.20**	**0.30**	**0.40**	**0.50**	**0.70**	**0.90**	**1.00**
3	10	0.5	F	0.047	0.082	0.108	0.250	0.462	0.642	0.776	0.870	0.918	0.924
			L	0.045	0.078	0.074	0.156	0.308	0.466	0.596	0.792	0.882	0.914
			T	0.050	0.426	0.546	0.640	0.752	0.798	0.846	0.858	0.898	0.928
		1.0	F	0.036	0.106	0.310	0.918	1	1	1	1	1	1
			L	0.034	0.078	0.138	0.418	0.806	0.882	0.916	0.930	0.986	1
			T	0.054	0.988	1	1	1	1	1	1	1	1
	30	0.5	F	0.056	0.104	0.240	0.890	1	1	1	1	1	1
			L	0.056	0.080	0.132	0.340	0.822	0.884	0.902	0.938	0.988	1
			T	0.058	0.940	0.990	1	1	1	1	1	1	1
		1.0	F	0.046	0.124	0.342	0.996	1	1	1	1	1	1
			L	0.046	0.086	0.160	0.432	0.872	0.870	0.878	0.956	0.984	1
			T	0.052	1	1	1	1	1	1	1	1	1
5	10	0.5	F	0.046	0.136	0.374	0.968	1	1	1	1	1	1
			L	0.042	0.100	0.180	0.486	0.812	0.904	0.956	0.965	0.986	1
			T	0.052	0.984	1	1	1	1	1	1	1	1
		1.0	F	0.052	0.144	0.359	0.988	1	1	1	1	1	1
			L	0.054	0.104	0.168	0.502	0.838	0.862	0.924	0.934	0.984	1
			T	0.056	1	1	1	1	1	1	1	1	1
	30	0.5	F	0.050	0.130	0.370	0.994	1	1	1	1	1	1
			L	0.052	0.076	0.178	0.442	0.802	0.89	0.892	0.922	0.980	1
			T	0.054	1	1	1	1	1	1	1	1	1
		1.0	F	0.044	0.124	0.352	0.996	1	1	1	1	1	1
			L	0.034	0.072	0.156	0.500	0.840	0.898	0.904	0.944	0.980	1
			T	0.050	1	1	1	1	1	1	1	1	1

Source: author computations; F: Fisher, L: Liptak, T: Tippett, τ: location shift, φ:combining function

In Table 2, the rejection rates of the tests when the number of variables is $V = 50$ are reported. Again the good performance of the tests under the null hypothesis ($p = 0.00$) is proved by the values of the estimated power. These values are usually not greater than $\alpha = 0.05$ even if sometimes they exceed the significance level, especially in the case of Tippet's combination. Nevertheless, when greater than 0.05, the rejection rates under H_0 are not far from α and then the tests are well approximated. Hence, this conclusion is valid regardless of the number of variables V.

Table 2 confirms also that the probability of right rejection of the null hypothesis of MANOVA by the combined permutation tests increases with the sample size n, with the number of samples S, with the shift parameter τ and with the proportion of true partial alternative hypotheses p. Another empirical evidence of the simulation study is that in general the power is greater with 100 variables than with 50 variables. This statement seems obvious thinking to the tendency of the power to one when the number of variables diverges in the two-sample problem proved by Pesarin and Salmaso (2010b). They focus on the relationship between power of the overall test and non-centrality parameter in the case 100% of the variables are under the alternative hypothesis. According to our results, the power of the multi-sample tests in the case $V = 100$ is much greater than in the case $V = 50$ only when the percentage of true partial alternative hypotheses is low, otherwise the difference seems not evident and always in the same direction. Hence, in our opinion, for the power behaviour, the proportion of true partial alternative hypothesis matters and it is more important than the absolute number of true partial alternatives. For instance, when $V = 50$ and $p = 0.40$, the number of true partial alternative hypothesis is 20, exactly as when $V = 100$ and $p = 0.20$. But in the former case, when $S = 3$, $n = 10$ and $\tau = 0.5$, the rejection rates of the tests based on the Fisher, Liptak and Tippett combination are 0.626, 0.490 and 0.636 respectively; instead in the latter case, under the same scenario, 0.250, 0.156 and 0.640 respectively. Hence, even if the number of true alternative hypotheses is the same, the power of the tests based on the Fisher and Liptak combinations is much lower when the proportion of true partial alternative hypotheses is smaller. Tippett represents an exception. Consider, under the same scenario, the case $V = 50$ and $p = 0.20$ (rejection rate 0.466) and $V = 100$ and $p = 0.10$ (rejection rate 0.546). Hence, with the same proportion p, the power increases with V only in the case of Tippett's combination.

In general, the case $V = 50$, confirms that the Liptak combination is the best choice only when $p = 1$ but in this situation the power of the other tests is very similar. In most of the considered settings, the Tippett combination is preferable because the power quickly tends to 1 as the proportion of true alternative hypotheses diverges. When $S = 3$, $n = 10$ and $\tau = 0.5$ this is the most powerful test up to $p = 0.40$. For larger values of p it becomes the less powerful test.

Table 2. Rejection rates of combined permutation tests for $V = 50$ and $\alpha = 0.05$.

S	n	τ	φ	Proportion of true partial alternative hypotheses (p)									
				0.00	**0.06**	**0.10**	**0.20**	**0.30**	**0.40**	**0.50**	**0.70**	**0.90**	**1.00**
3	10	0.5	F	0.050	0.082	0.096	0.230	0.414	0.626	0.766	0.870	0.924	0.886
			L	0.054	0.066	0.074	0.142	0.258	0.490	0.668	0.828	0.912	0.888
			T	0.057	0.268	0.316	0.466	0.556	0.636	0.726	0.768	0.818	0.810
		1.0	F	0.042	0.054	0.260	0.892	0.996	1	1	1	1	1
			L	0.038	0.068	0.120	0.446	0.812	0.938	0.950	0.986	0.988	1
			T	0.050	0.094	0.966	1	1	1	1	1	1	1
	30	0.5	F	0.052	0.100	0.230	0.824	1	1	1	1	1	1
			L	0.054	0.070	0.138	0.348	0.756	0.936	0.954	0.974	0.992	1
			T	0.056	0.876	0.960	0.994	0.998	0.998	1	1	0.998	1
		1.0	F	0.038	0.128	0.278	0.996	1	1	1	1	1	1
			L	0.040	0.092	0.140	0.424	0.846	0.938	0.948	0.974	0.980	1
			T	0.050	1	1		1	1	1	1	1	1
5	10	0.5	F	0.038	0.164	0.310	0.904	0.998	1	1	1	1	1
			L	0.044	0.132	0.162	0.408	0.798	0.944	0.952	0.970	0.988	1
			T	0.051	0.946	0.994	0.996	1	1	1	1	1	1
		1.0	F	0.048	0.182	0.174	0.978	1	1	1	1	1	1
			L	0.052	0.114	0.356	0.452	0.826	0.950	0.948	0.970	0.994	1
			T	0.054	1	1	1	1	1	1	1	1	1
	30	0.5	F	0.052	0.126	0.316	0.990	1	1	1	1	1	1
			L	0.048	0.076	0.156	0.458	0.836	0.940	0.956	0.976	0.996	1
			T	0.054	1	1	1	1	1	1	1	1	1
		1.0	F	0.051	0.136	0.348	0.986	1	1	1	1	1	1
			L	0.049	0.072	0.156	0.468	0.852	0.938	0.954	0.976	0.990	1
			T	0.053	1	1	1	1	1	1	1	1	1

Source: author’s computations; F: Fisher, L: Liptak, T: Tippett, τ: location shift, φ:combining function

5. Case Study about organizational well-being of University Workers

Organizational well-being is the first element that influences effectiveness, efficiency, productivity and development of a public organization. As part of objective 3 of the 2014-2016 Positive Action Plan proposed by the Equality Opportunities Office of the University of Ferrara (UNIFE), the Rector's Delegate for Equal Opportunities presented a project in order to promote the improvement of the working well-being of the administrative-technical staff. This project consists in the definition of interventions aimed at improving quality of working life based on findings deriving from empirical surveys.

A questionnaire was administered to a sample of 120 employees of UNIFE in order to assess the degree of work-related stress, to detect the opinions of employees with respect to the organization and the working environment and identify possible actions for the improvement of the general conditions of the public employees at UNIFE. One goal of the survey was also to test the existence of possible differences in organizational well-being among sub-groups of employees defined by gender and age.

The 120 respondents represent a random sample of the population of the technical-administrative staff. In order to test for the joint effect of gender and age on the organizational well-being at UNIFE, a simple random sample of 30 employees was selected from each of the following four groups:

- FU50: 50 years old or younger females,
- FO50: over 50 years old females,
- MU50: 50 years old or younger males,
- MO50: over 50 years old males.

The questionnaire, consisting of 79 questions, was administered to the respondents from the 4th to the 11th of December 2014. The questionnaire was designed by the Italian National Anti-Corruption Authority (ANAC) and the National Institute for Occupational Accident Insurance (INAIL) that decided to adopt a Likert scale, based on the first 6 integer values representing the level of agreement with respect to the 79 statements (1= not at all, …, 6=completely). It is worth noting that the permutation analysis of variance can be applied to numeric variables. The assumption of normality but not even that of continuity is required. Hence, it is a valid approach in the case of both interval and discrete scales. The results of the simulation study can be extended to testing problems for interval variables, and consequently applied to the case study. Even if, strictly speaking, the response variables in the considered application on organizational well-being are ordinal, it is common practice to treat them as interval data. In general, interval and discrete variables can be considered as the result of the discretization of continuous variables.

Furthermore, unlike the parametric approach, the permutation test does not require that a specific underlying family of distributions is known or assumed. The null permutation distribution of the test statistics can be determined regardless of whether the underlying distribution of the data is continuous or not. The 79 statements are reported in Appendix 1.

Let Y_{vg} be the random variable that represents the response concerning the v-th statement of an employee belonging to group g, with $v = 1,2, \dots ,79$ and $g \in G = \{FU50, FO50, MU50, FO50\}$. The testing problem can be represented by the following hypotheses:

$$H_0: \bigcap_{v=1}^{79} \left[Y_{v,FU50} =^d Y_{v,FO50} =^d Y_{v,MU50} =^d Y_{v,MO50} \right]$$

vs

$$H_1: \bigcup_{v=1}^{79} \left[\exists g', g'' \in G \text{ s.t. } Y_{v,g'} \neq^d Y_{v,g''} \right]$$

The significance level is $\alpha = 0.05$. According to the simulation study, the most suitable testing method seems to be the combined permutation test based on the Tippett combining function. The application of this test provides a p-value of 0.755, much greater than α. Hence the null hypothesis cannot be rejected. At the significance level 0.05, there is no empirical evidence to reject the null hypothesis of no difference of the organizational well-being between groups in favor of the hypothesis that the organizational well-being of the groups is not the same. In other words, we cannot conclude that there is a significance effect of gender and age on the employees' well-being. The analysis was carried out by the authors by creating specific R scripts for the implementation of the methodology.

It is worth noting that the final p-value of the combined test is invariant with respect to the combination strategy. In other words, if we perform a two-level combination, i.e. the first within-domain combination of partial tests and the second combination with respect to the domains, the final result is the same as obtained by permuting the partial tests all together at the same time (see Pesarin, 2001). If we had significance in the overall test, it would be useful to identify the partial tests that contribute to the overall significance. This can be done with a suitable adjustment of the p-values of the partial tests for controlling the Family Wise Error rate and avoiding the inflation of the type I error of the final combined test.

In this case, an interesting two-stage combination strategy could be of interest, because the questionnaire is divided into sections corresponding to partial aspects of organizational well-being. Each aspect corresponds to a set of questions and consequently to a domain of variables (construct). In the case of significance of the overall combined test, the analysis of the adjusted p-values of the partial combined tests related to the constructs would make sense. Unfortunately, the overall null hypothesis is not rejected.

This result proves that, in the University of Ferrara, the organizational well-being of the employees in terms of risks, working environment, respect, relationship with colleagues and office manager, transparency, motivation, etc. is not affected by age and gender. It could be considered as evidence of gender-age equality within the organization.

6. Conclusions

The purpose of the work is to deepen the study of the power behaviour of combined permutation tests for MANOVA problems with big data. The assessment of the convergence rate of the power to one as the proportion of variables under the alternative hypothesis increases and a comparison between the three most commonly used members within this family of tests represent the main scientific added value of the paper.

These nonparametric multi-sample location tests are well approximated, consistent, unbiased and powerful also for small sample sizes. The power is also an increasing function of the number of samples and of the number of variables of the dataset. The asymptotic behaviour of the tests when the number of variables diverges was studied and the simulations proved that the proportion of true partial alternative hypotheses is more important than the absolute number of variables of the dataset in explaining the increase of power. The test based on the Tippett combination represents an exception to this general rule.

This test seems to be much more powerful than the others when the proportion of true partial alternative hypotheses is not large but competitive also when the proportions are close to one. This is the only condition in which the test based on Liptak combination is competitive but, for small proportions of true alternatives, this test is by far the least powerful.

Definitely, it seems that, among the distribution free solutions to the multivariate analysis of variance in the family of combined permutation tests, the method based on the Tippet combination is in general preferable, especially if there are no preventive information about the possible percentage of variables (or marginal distributions) under the alternative hypothesis. Instead of the Tippett combination, the Fisher rule can be applied when the percentage is close to 100%. The Liptak combination seems to be non-convenient in general.

This methodological tool is an important and useful solution of testing problems for big data, especially when the number of variables is very large and the sample sizes are small. The usefulness and the effectiveness of the method is confirmed by the application to the case study concerning the survey on the organizational well-being at the University of Ferrara discussed in the paper.

Acknowledgments

The authors wish to thank the anonymous reviewers who have contributed significantly to the quality of the paper with their comments and suggestions.

The work was partially supported by the University of Ferrara, which funded the FIR project "Measuring and assessing social sustainability, inclusion and accessibility at university: methods and applications".

References

Anderson, M. J., (2001). A new method for non-parametric multivariate analysis of variance. *Austral ecology*, 26(1), pp. 32–46.

Arboretti, R., Bonnini, S., (2008). Moment-based multivariate permutation tests for ordinal categorical data. *Journal of Nonparametric Statistics*, 20(5), pp. 383–393.

Arboretti, R., Bonnini, S., (2009). Some new results on univariate and multivariate permutation tests for ordinal categorical variables under restricted alternatives. *Statistical Methods and Applications: Journal of the Italian Statistical Society*, 18(2), pp. 221–236.

Arboretti, R., Ceccato, R., Corain, L., Ronchi, F. and Salmaso, L., (2018). Multivariate small sample tests for two-way designs with applications to industrial statistics. *Statistical Papers*, 59(4), pp. 1483–1503.

Baro, E., Degoul, S., Beuscart, R. and Chazard, E., (2015). *Toward a literature-driven definition of big data in healthcare.* BioMed research international (https://doi.org/10.1155/2015/639021).

Bonnini, S., And Melak Assegie, G., (2019). *Permutation multivariate tests for treatment effect: theory and recent developments.* In SUSAN SSACAB 2019, pp. 30–30. The Biostatistics Research Unit of the South African Medical Research Council.

Bonnini, S., (2014). Testing for heterogeneity with categorical data: permutation solution versus bootstrap method. *Communications in Statistics: Theory and Methods*, 43(4), pp. 906–917.

Bonnini, S., (2016). Multivariate approach for comparative evaluations of customer satisfaction with application to transport services. *Communications in Statistics: Simulation and Computation*, 45(5), pp. 1554–1568.

Bonnini, S., Corain, L., Marozzi, M. and Salmaso, L., (2014). *Nonparametric hypothesis testing: rank and permutation methods with applications in R. John Wiley & Sons.*

Bonnini, S., Prodi, N., Salmaso, L., Visentin, C., (2014). Permutation approaches for stochastic ordering. *Communications in Statistics: Theory and Methods*, 43(10-12), pp. 2227–2235.

Clarke, K.R., (1993). Non-parametric multivariate analyses of changes in community structure. *Australian journal of ecology*, 18(1), pp.117–143.

Farcomeni, A. and Greco, L., (2016). Robust methods for data reduction. CRC press.

Finch, W.H., (2016). Comparison of multivariate means across groups with ordinal dependent variables: a Monte Carlo simulation study. *Frontiers in Applied Mathematics and Statistics*, 2, p. 2.

Hotelling, H., (1992). The generalization of Student's ratio. I*n Breakthroughs in statistics*, (pp. 54-65). Springer, New York, NY.

Johnson, R., (1997). Wichern. D., (2007). Applied multivariate statistical analysis. Prentice-Hall: London.

Legendre, P. and Anderson, M. J., (1999). Distance-based redundancy analysis: testing multispecies responses in multifactorial ecological experiments. *Ecological monographs*, 69(1), pp.1–24.

Mantel, N., Valand, R. S., (1970). A technique of nonparametric multivariate analysis. *Biometrics*, pp. 547-558.

McArdle, B. H., Anderson, M. J., 2001. Fitting multivariate models to community data: a comment on distance-based redundancy analysis. *Ecology*, 82(1), pp. 290–297.

Mielke Jr, P. W., Berry, K. J., (1999). Multivariate tests for correlated data in completely randomized designs. *Journal of Educational and Behavioral Statistics*, 24(2), pp. 109–131.

Mielke Jr, P. W., Berry, K. J., Johnson, E. S., (1976). Multi-response permutation procedures for a priori classifications. *Communications in Statistics: Theory and Methods*, 5(14), pp. 1409–1424.

Özköse, H., Arı, E. S. and Gencer, C., (2015). Yesterday, today and tomorrow of big data. *Procedia-Social and Behavioral Sciences*, 195, pp. 1042–1050.

Pesarin, F., (2001). *Multivariate permutation tests: with applications in biostatistics*, Vol. 240. Wiley: Chichester.

Pesarin, F., Salmaso, L., (2010a). *Permutation tests for complex data: theory, applications and software*. John Wiley & Sons: Chichester.

Pesarin, F., Salmaso, L., (2010b). Finite-sample consistency of combination-based permutation tests with application to repeated measures designs. *Journal of Nonparaetric Statistics*, 22(5), pp. 669–684.

Pillai, K. S., (1955). Some new test criteria in multivariate analysis. *The Annals of Mathematical Statistics*, pp. 117–121.

Pillar, V., (2013). How accurate and powerful are randomization tests in multivariate analysis of variance?. *Community Ecology*, 14(2), pp. 153–163.

Pillar, V.D.P., Orlóci, L., (1996). On randomization testing in vegetation science: multifactor comparisons of relevé groups. Journal of Vegetation Science, 7(4), pp. 585–592.

Polko-Zajac, D., (2019). On permutation location-scale tests. *Statistics in Transition*, 20(4), pp. 153-166.

Polko-Zajac, D., (2020). A comparative study on the power of parametric and permutation tests for a multidimensional and two-sample location problem. *Argumenta Oeconomica Cracoviensia*, 2(23), pp. 69–79

Wilks, S. S., (1932). Certain generalizations in the analysis of variance. *Biometrika*, pp. 471–494.

Appendix 1

Code	Statement
A.01	My working place is safe
A.02	I have been informed about the risks connected to my job
A.03	I am satisfied about the environment of my working place
A.04	I have suffered harassment
A.05	My dignity has been harmed at work
A.06	At work the smoking ban is respected
A.07	I usually take enough breaks
A.08	I can work hard
A.09	I am not comfortable when I am working
A.10	The colleagues are not polite with me
A.11	I am allowed to take a break when I wish
A.12	I don't have the chance to take enough breaks
B.10	At work I have suffered bullying
B.01	In the workplace I am respected in my trade union membership
B.02	In the workplace I am respected in my political orientation
B.03	In the workplace I am respected in my religious faith
B.04	My gender identity is an obstacle to my enhancement at work
B.05	In the workplace I am respected in my ethnicity and race
B.06	In the workplace I am respected in relation to my mother tongue
B.07	My age is an obstacle to my enhancement at work
B.08	In the workplace I am respected in relation to my mother tongue
C.01	The workload is assigned with equity
C.02	The responsibilities are assigned with equity
C.03	My salary is proportional to the commitment
C.04	The pay is differentiated according to quantity and quality of work
C.05	My manager makes work decisions impartially
D.01	At UNIFE the path of professional development of each employee is well defined and clear
D.02	At UNIFE the career opportunities depend on merit
D.03	UNIFE gives the possibility to develop skills and aptitudes of individuals in relation to the requirements of the different roles
D.04	My current role is appropriate to my professional profile
D.05	I am satisfied with my professional path within UNIFE
E.01	I know what is expected of my work
E.02	I have the skills to do my job
E.03	I have the resources and tools to do my job
E.04	I have an adequate level of autonomy in my work
E.05	My work gives me a sense of personal fulfilment
E.06	I know how to do my job
E.07	I understand what is expected of me at work
E.08	I have freedom of choice in deciding how to do my job
E.09	I have unattainable deadlines

Code	Statement
E.10	I have to work very hard
E.11	I have a say in deciding how fast I can do my job
E.12	I’m getting pressure to work overtime
E.13	I have freedom of choice in deciding what to do at work
E.14	I have to do my job very quickly
E.15	I have deadlines impossible to meet
E.16	I have a say in how to do my job
E.17	My working hours can be flexible
E.18	Job requests made to me by various people/offices are difficult to combine
F.01	I feel part of a team
F.02	I help colleagues even if it’s not my job
F.03	I am esteemed and treated with respect by colleagues
F.04	In my group, those who have information make it available to everyone
F.05	The organization pushes to work in a group and to collaborate
F.06	If the job becomes difficult, I can count on the help of my colleagues
F.07	At work my colleagues show me the respect I deserve
F.08	I receive support information that helps me in my work
F.09	There are frictions or conflicts between colleagues
F.10	My colleagues give me the help and support I need
F.11	Colleagues are willing to listen to my work problems
G.01	My organization invests in people, including through adequate training
G.02	The rules of conduct are clearly defined
G.03	Organisational tasks and roles are well defined
G.04	The circulation of information within the organisation is appropriate
G.05	My organisation promotes measures to reconcile working time and life time
G.06	I have clear duties and responsibilities
G.07	I must neglect some tasks because I have too much to do
G.08	I know the goals of my department/office
G.09	Staff are always consulted on changes in work
G.10	I’m supported in emotionally challenging jobs
G.11	Workplace relations are strained
H.01	I am proud when I tell someone that I work at UNIFE
H.02	I am proud when UNIFE achieves good results
H.03	I am sorry if someone has a bad opinion of UNIFE
H.04	Values and behaviours at UNIFE are similar to mine
H.05	If possible, I would change company
I.01	Relative and friends think that UNIFE is important for the collectivity
I.02	Students think that UNIFE is important for the collectivity
I.03	People think that UNIFE is important for the collectivity