Quaderni di Statistica
Vol. 12, 2010

Nonparametric multivariate ranking methods for global performance indexes

Rosa Arboretti Giancristofaro
Department of Territory and Agri-Forestal Systems, University of Padova
E-mail: rosa.arboretti@unipd.it

Livio Corain
Department of Management and Engineering, University of Padova
E-mail: livio.corain@unipd.it

Daniele Gomiero Federico Mattiello
Department of Statistics, University of Padova
E-mail: daniele.gomiero@yahoo.it ; federico.mattiello@gmail.com

Summary: The proper definition of a global performance index is a challenging topic, especially in the field of New Product Development and Education. Very often in this context, the research aim is focused on evaluating the performances of treatments (products, services, etc.) from a multivariate point of view, that is, in connection with more than one aspect and/or under several conditions. Therefore, the main goal of statistical data analysis consists in the calculation of a proper index to obtain a global performance evaluation of the treatments under investigation. The purpose of this work is to present an innovative nonparametric method for ranking of treatments, with reference to the analysis of variance layout, using a suitable global performance index, and to critically compare two challenging indexes that can be used in complex situations whereas parametric procedures can not be reliably employed. The goal of this paper is to find which procedure is more reliable. In particular, the procedures are tested by varying experimental conditions such as the number of variable and the distributions of random errors.

Keywords: Global ranking, Multivariate inference, Pairwise comparisons.

1. Introduction and motivation

Applied research problems are often related to datasets observed over more units (subjects, samples of product unit, etc.), with reference to several variables (evaluations, product performances, etc.), with the aim of studying the relationships between these variables and a factor of interest under investigation (a given firm's feature, product, etc.). In this framework the main goal is to compare the factor levels, with respect to all variables, in order to rank them and hence to find out the "best" one.

Recently, in the literature there's a growing interest on the topic of multivariate ranking methods of treatments. Despite the fact that the literature of multiple comparison methods addresses the problem of ranking a set of treatment groups from worst to best (Westfall *et al.*, 1999), no clear indication is provided on how dealing with the information from pairwise multiple comparisons, especially in case of blocking and/or multivariate response variable. Since the seminal paper by Bonnini *et al.* (2006), several other methodological contributions have been presented in the literature (Corain and Salmaso, 2007; Arboretti *et al.*, 2008; Bonnini *et al.*, 2009).

The ranking and selection approach in multiple decision theory, as it can be seen in Gupta and Panchapakesan (2002) (which contains an extensive discussion on the whole theory), provides some hints on the topic, but essentially for univariate problems and under assumption of normality. Moreover, although this book deals with a great number of available procedures, it is more focussed on theoretical aspects such as defining ranking rules that respect a given probability of correct selection, or providing formulae for choosing the minimum sample size such that this probability is attained, rather than providing practical rules that can be directly used in real situations.

This problem is not only of theoretical interest but also it has a recognized relevance. In fact, especially for industrial research, a global ranking in terms of performance of all investigated products/prototypes is a very natural goal (Bonnini *et al.*, 2006). When performance evaluation takes into account more than one aspect, the problem can be complicated and some methodological and practical issues arise: standardization, mul-

tivariate structure of data, accuracy of partial indicators, distance with respect to target (highest satisfaction level), stratification in presence of confounding factors (Fayers and Hand, 2002).

As confirmation of the genuine interest on the topic by practitioners, it is worth noting that in 2008 an international industrial organization called AISE has formally incorporated such a methods as official standard for industrial research on house cleaning products (AISE, 2009). AISE is the international Association for Soaps, Detergents and Maintenance Products. It is the official representative body of this industry in Europe. Its membership totals 37 national associations in 42 countries, covering about 900 companies ranging from small and medium-sized enterprises to large multinationals active both in the consumer goods market and the industrial and institutional domains (AISE, 2010).

The present paper is organized as follows: section 2 provides the formalization of the problem, defines the theoretical background and describes the concept of ranking parameter along with its use; section 3 is devoted to the description of the algorithm used to obtain the multivariate ranking; in section 4 we present details on the simulation study used for comparing the proposed procedures; finally, section 5 contains final discussion and some conclusions, including some purposes for future researches.

2. Formalisation of the problem

In order to introduce how an inferential approach for ranking of multivariate populations can be developed, let $\mathbf{Y}_{ik}$ be the multivariate numeric variable related to the p-variate response of any experiment of interest and let us assume, without loss of generality, that high values of each marginal univariate component corresponds to better performances and therefore to a higher degree of preference. The experimental design of interest is defined by the comparison of C groups or treatments with respect to p different variables where n replications of a single experiment are performed by a random assignment of statistical units to treatments (or groups). The C-group multivariate statistical model (with fixed ef-

fects) can be represented as follows:

$$\mathbf{Y}_{ik} = \boldsymbol{\mu}_i + \boldsymbol{\varepsilon}_{ik}, \;\; \boldsymbol{\varepsilon}_{ik} \sim \text{IID}(0, \boldsymbol{\Sigma}); \quad i = 1, ..., C; \; k = 1, \ldots, n \; ; \quad (1)$$

where, in the case of a balanced design, n is equal to the number of replications, index i is related to treatments with p-variate mean effect equal to $\boldsymbol{\mu}_i$, index k is related to replications and $\boldsymbol{\varepsilon}_{ik}$ is a p-variate random term of experimental errors with zero mean and variance-covariance matrix $\boldsymbol{\Sigma}$.

In order to determine whether the groups/treatments are equivalent against the alternative that they are different, we introduce the following hypothesis testing layout. At first a multivariate global C-sample hypothesis H_0^G is considered. If this is rejected, inference will concern the multivariate pairwise comparisons H_0^{ih}, and possibly on univariate pairwise comparisons $H_{0|j}^{ih}$. More formally,

$$\begin{cases} H_0^G : \boldsymbol{\mu}_1 = \cdots = \boldsymbol{\mu}_C \\ H_1^G : \exists\, i, h \;\big|\; \boldsymbol{\mu}_i \neq \boldsymbol{\mu}_h \end{cases} \quad (2a)$$

$$\text{then, if } H_0^G \text{ is rejected} \quad \begin{cases} H_0^{ih} : \boldsymbol{\mu}_i = \boldsymbol{\mu}_h \\ H_1^{ih} : \boldsymbol{\mu}_i \neq \boldsymbol{\mu}_h \end{cases} \quad (2b)$$

$$\text{then, if } H_0^{ih} \text{ is rejected} \quad \begin{cases} H_{0|j}^{ih} : \mu_{ij} = \mu_{hj} \\ H_{1|j}^{ih} : \mu_{ij} \neq \mu_{hj} \end{cases} \quad (2c)$$

$$i, h = 1, \ldots, C, \; i \neq h, \; j = 1, \ldots, p \, .$$

Since the focus of this work is not only on hypothesis testing, but also on defining and estimating a suitable indicator to quantify the relative preference of each treatment in comparison with each other (in order to rank them), let us consider a so-called "ranking parameter" $\theta_i = f_i(\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_C)$, $i = 1, \ldots C$, such that the rank transformation of θ_i may be able to provide a meaningful ranking of the i-th treatment from a multivariate point of view (the concept of "ranking parameter" can be found in Gupta and Panchapakesan, 2002).

It is worth noting that the choice of the functions $f_i(.)$ is particularly sensitive, in fact it represents the way in which the data dimensionality is reduced, hence in general:

- we cannot think on an optimal solution because it depends on the unknown underlying data structure;
- goodness of $f_i(.)$ depends on the underlying metric, so that geometrical functions can be used only with continuous random variables while in case of categorical r.v.s it is more appropriate to use a more robust approach (such as goodness-of-fit functions).

  Examples of ranking parameters are:

- Euclidean distance: $\theta_i^{\text{dist}} = \|\boldsymbol{\mu}_i - \boldsymbol{\mu}_0\|$;
- squared Mahalanobis distance: $\theta_i^{\text{Mah}} = (\boldsymbol{\mu}_i - \boldsymbol{\mu}_0)^{\top} \cdot \boldsymbol{\Sigma}^{-1} \cdot (\boldsymbol{\mu}_i - \boldsymbol{\mu}_0)$;

where $\boldsymbol{\mu}_0$ is a known reference p-dimensional point, for example the minimum or maximum value that can be reached by the response variable. Mahalanobis distance differs from Euclidean distance in that it takes into account the correlations of the data-set and is scale-invariant, i.e. it is not dependent on the scale of measurements. For this reasons, the Mahalanobis distance is often preferred with respect to the Euclidean distance. Note that we are implicitly assuming that all response variables are defined in the same metric and this is true in many real cases of interest.

By combination of the p-values directly related to the set of original univariate testing procedures (see expression 2c), a robust and even more informative ranking parameter can be defined:

- NPC score: $\theta_i^{\text{NPC}} = -2 \sum_{j=1}^{p} \sum_{\substack{h=1 \\ h \neq i}}^{C} \log \mathrm{p}_{ih|j}$;

where $\mathrm{p}_{ih|j}$ is a p-value suitable for testing the hypothesis $H_{0|j}^{ih}$ and here are calculated using the unknown population means (actual parameters). Note that the NPC score (NonParametric Combination) is actually the so-called nonparametric Fisher combining function, often used to derive multivariate testing procedures (see Pesarin and Salmaso, 2010). Note that the NPC score depends on the test statistics involved in it and it is a function not only of all the true means but also of the unknown dependence structure of the multivariate random errors. In order to make

the NPC score more informative with respect to our goal of ranking the multivariate treatments, we can take account of directional type p-values, namely those that are suitable for testing the hypotheses:

$$\begin{cases} H_{0|j}^{ih} : \ \mu_{ij} \leq \mu_{hj} \\ H_{1|j}^{ih} : \ \mu_{ij} > \mu_{hj} \end{cases} \qquad (3)$$
$$i \neq h, \quad i,h = 1,\ldots,C, \quad j = 1,\ldots,p.$$

Note that, similarly to the distance-based ranking parameters, the NPC score has the following characteristics: (i) it takes values greater (or equal) to zero and tends to take lower values when the hypothesis of equality of treatments is true; (ii) conversely, it tends to take large values under the alternative hypothesis of difference between treatments.

The fact of using directional p-values helps us to make the NPC score more suitable to our problems and to better discriminate treatments in order to obtain a ranking of them. In order to achieve the objective of finding out a ranking of multivariate treatments, let us now rewrite the inferential problem in terms of ranking parameters:

$$\begin{cases} H_0^G : \boldsymbol{\mu}_1 = \cdots = \boldsymbol{\mu}_C \\ H_1^G : \exists i,h \,\big|\, \boldsymbol{\mu}_i \neq \boldsymbol{\mu}_h \end{cases} \rightsquigarrow \quad \begin{cases} {}^{\theta}H_0^G : \theta_1 = \cdots = \theta_C \\ {}^{\theta}H_1^G : \exists i,h \,\big|\, \theta_i \neq \theta_h \end{cases} \qquad (4a)$$

$$\text{then, if } {}^{\theta}H_0^G \text{ is rejected} \qquad \begin{cases} {}^{\theta}H_0^{ih} : \theta_i = \theta_h \\ {}^{\theta}H_1^{ih} : \theta_i \neq \theta_h \end{cases} \qquad (4b)$$
$$i \neq h, \ \ i,h = 1,\ldots,C \ ;$$

where for the univariate test statistics in the NPC methods we use directional p-values as it is described in section 2.

Note that the two approaches of testing of hypotheses can not be considered equivalent. However in this context we are more interested in estimation and ranking than in hypothesis testing, provided that in this conversion the "lack of information" is as little as possible. In fact, if the global null hypothesis on the original parameter $\boldsymbol{\mu}_i$ is true, then the global null hypothesis on the ranking parameters is also true, but in general the *viceversa* does not always hold. This is due to the reduction of

dimensionality operated by the synthesis functions $f_i(.)$s in which two treatments could differ in opposite direction in two different variables, hiding these differences on the global testing side. Nevertheless p-values calculation that will be subsequently described is based on the (pairwise) hypothesis testing 3, in the sense that they could be used to perform those tests.

Thus the multivariate inferential problem of interest can be viewed as a simultaneously interval estimation procedure on the differences $\theta_{ih}^T = \theta_i^T - \theta_h^T \quad i \neq h \quad i,h = 1,\dots C$; where T is the type of the ranking parameter we decide to adopt. As will be discussed afterwards, the problem becomes only apparently univariate because the ranking parameter estimator depends on the multivariate distribution of the error components ε's. This consideration applies even more to the NPC score ranking parameter, where a single θ_i depends on all the comparisons involving the i-th treatment in all variables at the same time. Depending on the assumptions made on the random errors, the distribution of ranking parameter estimators can be derived in a parametric or in a nonparametric way. For example, when assuming the multivariate normal distribution for random errors, the first score-statistic has an exact $\mathcal{F}$-type distribution, namely:

$$\widehat{\theta}_i^{\text{Mah}} = \widehat{\boldsymbol{\mu}}_i^{\top} \cdot \widehat{\boldsymbol{\Sigma}}_i^{-1} \cdot \widehat{\boldsymbol{\mu}}_i \quad \sim \quad \mathcal{F}_{p,\, n-p} \; ; \tag{5}$$

with reference to the second score-statistic, i.e. NPC-score, in general the following asymptotic result holds:

$$\widehat{\theta}_i^{\text{NPC}} = -2 \sum_{j=1}^{p} \sum_{\substack{h=1 \\ h \neq i}}^{C} \log \widehat{\mathrm{p}}_{ih|j} \; \xrightarrow{d} \; a \cdot \chi_g^2 \; ; \tag{6}$$

where n is the sample size, p is the number of variables, $\widehat{\boldsymbol{\Sigma}}_i$ is the $p \times p$ sample variance matrix calculated for every treatment and $\widehat{\mathrm{p}}_{ih|j}$ are p-values of the statistics for the (ih)-th comparison in the j-th variable calculated as it is described in subsection 3.1 on page 88.

It is noteworthy that actually we are interested in differences *between* pairs of ranking parameter estimators ($\widehat{\theta}_{ih}^T$), rather than individual ones (as it is implicit in formulae above).

In the case of Mahalanobis score, for the estimator of a pairwise difference we may easily refer to the Hotelling's T^2 distribution, while in the case of the NPC score, the asymptotic distribution is hard to find because of the coefficient a that multiplies the χ^2_g, which is a measure of dependence between the p-values involved into the statistic and could be different between different treatments; moreover it is related with the degrees of freedom g.

Furthermore the parametric approach presents a number of drawbacks that have to be taken into account.

1. When keeping the sample size fixed, increasing of dimensionality (number of variables) results in a loss of degrees of freedom, hence the estimation procedures may become inaccurate.

2. Under non normal errors, inferential achievements are valid solutions only asymptotically, so for finite samples (sample sizes are very small indeed, in context of this work) the approximation accuracy has to be carefully considered.

3. If the observations cannot be reasonably assumed as a sample from a multivariate normal distribution (as it is in many real situations), results can be inconsistent.

Conversely, the nonparametric resampling-based approach offers a number of advantages.

1. It is a robust solution, with respect to the true underlying distribution of response variables.

2. The dependence structure of the response variables is implicitly captured, so there is no need to estimate any dependence coefficient or to assume any dependence model.

3. It can be used with (virtually) arbitrary complicated indicator (i.e. ranking parameters).

3. A nonparametric multivariate ranking algorithm

In this section we present a general nonparametric resampling-based algorithm devoted to obtain a ranking of several multivariate population of interest by means of point and interval estimation of a global performance index. The proposed algorithm consists in two main steps: (i) score definition; (ii) estimation of the confidence intervals for pairwise differences of scores.

3.1. Step 1: defining scores

The step one of the proposed algorithm is described by the following stages:

1. Choose a suitable score-statistic (i.e. ranking parameter estimator) that summarise the relative position of each treatment (on the metric of the statistic). This results in a C-dimensional vector of scores.

2. Construct the confidence intervals for the pairwise differences of ranking parameter estimators $\widehat{\theta}_{ih}$, as described in the subsequent section.

3. Use these scores to test the $C \times (C-1)/2$ pairwise comparisons hypotheses 4b (see the previous section). This results in a $C \times C$ zero-one matrix for the rank assignment rule (described in subsection 3.2).

Because of their complexity, from a parametric point of view, the distribution of these ranking parameter estimators will be obtained via bootstrap and permutation resampling. The first one is the squared version of the Mahalanobis distance from the origin, here the observations are considered as a random sample from a p-variate distribution with the same variance-covariance matrix for every group (positive definite). This score is:

$$\widehat{\theta}_i^{\text{Mah}} = \overline{y}_i^{\top} \cdot \widehat{\mathbf{\Sigma}}^{-1} \cdot \overline{y}_i \,, \quad i = 1, \ldots, C \tag{7}$$

Where $\overline{y}_i$ is the vector of sample means for the i-th treatments. Note that because homoscedasticity assumption, $\boldsymbol{\Sigma}$ is estimated by her sample version using all the $n \times C$ observations.

In order to calculate the second ranking parameter estimator, called *NPC* (NonParametric Combination), let us consider the j-th response variable. We used two different approaches to obtain (estimated) p-values that have to be combined with the Fisher's combining function.

In the first approach we have considered parametric p-values (so the reference null distribution is the asymptotic one), performing the following steps:

- the statistics of a two-sample t-test is calculated for each of the $C \times (C-1)$ pairwise comparisons between two treatments [1], formally:

$$T_{ih|j} = \frac{\overline{y}_{i|j} - \overline{y}_{h|j}}{\sqrt{2\widehat{\boldsymbol{\sigma}}_j^2/n}}, \quad i \neq h, \quad i,h = 1,\ldots,C\ ; \tag{8}$$

  where $\overline{y}_{i|j}$ is the sample mean of the i-th treatment in the j-th variable and $\widehat{\boldsymbol{\sigma}}_j^2$ is the residual deviance resulted from fitting a one-way ANOVA model on that variable. Note that the numerator of the statistic is taken with its sign because the reference hypothesis is the 3 on page 84. This statistic has distribution $t_{C\cdot(n-1)}$ if $H_{0|j}^{ih}$ is true and errors are normally distributed;

- p-values are then calculated with $\mathbb{P}\left[\mathrm{T} \geqslant T_{ih|j}\right] = \widehat{\mathrm{p}}_{ih|j}$; where T has distribution $t_{C\cdot(n-1)}$.

In the second approach we have considered permutation p-values (hence the reference null distribution is the permutation one), performing the following steps:

- the statistic to be calculated for each of the $C \times (C-1)$ pairwise comparisons between two treatments is:

$$T_{ih|j} = \overline{y}_{i|j} - \overline{y}_{h|j}, \quad i \neq h, \quad i,h = 1,\ldots,C\ ; \tag{9}$$

[1] Note that the comparisons with indexes $(K+1,\ldots,2K)$ are equal to the comparisons with indexes $(1,\ldots,K)$ with changed signs. Here we calculate all the $2K$ comparisons in order to test each group against each other.

where $\overline{y}_{i|j}$ is the sample mean of the i-th treatment in the j-th variable

- permutation p-values are then calculated with

$$\frac{1}{B}\,\#\left({}^{b}T_{ih|j} \geq T_{ih|j}\right) = \frac{1}{B}\,\#\left[\left({}^{b}\overline{y}_{i|j} - {}^{b}\overline{y}_{h|j}\right) \geq \left(\overline{y}_{i|j} - \overline{y}_{h|j}\right)\right] \;; \tag{10}$$

where the superscript b indicate that the statistics is calculated using the b-th permuted sample, $\#(.)$ is the function that count the number of elements of the set that satisfies the condition (here the set is the B values of the statistic) and the statistic without superscript is the observed one.

Thus each treatment is matched against all others as if we were to test the system of hypotheses 3 (see the previous section). These steps has to be repeated for every variable obtaining a set of $p \times C \times (C-1)$ p-values.

Finally the global score for the i-th treatment is calculated with:

$$\widehat{\theta}_i^{\mathrm{NPC}} = -2\sum_{j=1}^{p}\sum_{\substack{h=1\\h\neq i}}^{C} \log \widehat{\mathrm{p}}_{ih|j} \;; \tag{11}$$

where we use as p-value combination the Fisher's combining function. It is noteworthy that this combining function is nonparametric with respect to the underlying dependence structure among p-values, since all kinds of monotonic dependencies are implicitly captured. The distribution of this statistics would be $\chi^2_{2p(C-1)}$ if all partial hypotheses involved are true and the terms of the summation are independent, but this is not the case. The p-values can not be considered independent because of the common denominator in the parametric p-value calculation, and because of the linear relation between $T_{i|j}$s, in the permutation p-value calculation.[2]

Even under H_0^G all we can say about this distribution is that its c.d.f. has the following property:

$$\mathbb{P}\left[\theta_i^{\mathrm{NPC}} \leq x\right] \in \left(\mathbb{P}\left[\mathrm{V} \leq x\right] \,,\, \mathbb{P}\left[\mathrm{W} \leq x\right]\right), \quad \forall x \in (0, +\infty) \tag{12}$$
$$\text{where} \quad \mathrm{V} \sim {}_{2p(C-1)}\cdot\chi^2_1 \,, \quad \mathrm{W} \sim \chi^2_{2p(C-1)};$$

[2] In fact if we have 3 groups then we get: $T_{12|j} = T_{13|j} - T_{23|j}$.

Furthermore, under H_1^{ih} the distribution is complicated by the dependence structure among the terms, another motivation supporting the bootstrap or the permutation approach.

3.2. Step 2: confidence intervals for pairwise differences of scores

After obtaining the bootstrap (or permutation) distributions of the considered ranking parameter estimators, in order to test the pairwise hypotheses, we decided to construct the $K = C \times (C-1)/2$ simultaneous confidence intervals for pairwise differences of scores. Hence the α (significance level) has been corrected for multiplicity with Bonferroni's method, i.e. $\alpha' = \alpha/K$. Then, after this, a one was associated to all comparisons for which the confidence interval contains the origin and a zero elsewhere, producing the $C \times C$ zero-one matrix for the ranking rule (for details see subsection 3.2).

It is worth to describe here how the issue of constructing confidence intervals just mentioned has been solved. The simplest way to build these confidence intervals is to directly use the bootstrap or permutation distribution, taking the sample $\alpha'/2$ and $1-\alpha'/2$ quantiles as estimates of true quantiles, in this way a proportion of α' observations are excluded from the interval and the confidence level is theoretically satisfied. Problems arise when α' is too small with respect to the sample size, as it is in our case (even though there are $B = 1000$ bootstrap or permutation replications), because estimates become inaccurate due to the small probability associated with the binomial distribution of the estimators (it is based on the empirical c.d.f. statistics).

To get rid of this limitations and following the ideas from Hinkley (1975) we decided to use the *unconditional* version of the Box-Cox's transformation (that is, there is no regression model), appeared in Box and Cox (1964) which expression is:

$$z_\lambda := \begin{cases} \dfrac{y^\lambda - 1}{\lambda}, & \text{if} \quad \lambda \neq 0 \\ \log y\ , & \text{if} \quad \lambda = 0 \end{cases} ; \tag{13}$$

and it is monotonic for every fixed λ, which is found with a numerical

maximisation. From interval estimates of θ_{ih} (or equivalently observed p-value related to H_0^{ih}) it is desirable to define a suitable algorithm able to estimate the multivariate ranking of the C treatments. In fact, only in a few cases when all differences are declared significant, it would be easy to find out a meaningful ranking but, since a sort of transitive property of significant differences obviously does not exist, we need a general rule able to assign a ranking to the multivariate treatments.

All the procedures that will be presented here have a common outcome: a zero-one $C \times C$ matrix containing the results of the pairwise comparisons. The (i, h)-th cell of this matrix, with $h > i$, takes the value "one" if the (i, h)-th pairwise null hypothesis is not rejected and "zero" elsewhere (hence if the two treatments can not be considered as equal); "one" where $i = h$ (every treatment is always equal to itself); while (i, h)-th cells with $h < i$ can be considered as *N.A.*s (Not Available values). Note that treatments have been ordered from the highest ("best") to the lowest ("worst") according to the point estimates ($\widehat{\theta}_i$s), before calculating pairwise comparisons (and hence before constructing the matrix), thus the first row contains the comparisons between the best treatment against each other, and so on.
Starting from this matrix the ranking rule can be described as follows:

1. row 1 is multiplied by 1, so the rank 1 is assigned to all treatments that are non-significantly different from (1) ("best"), including (1) itself;

2. row 2 is multiplied by 2, so the rank 2 is assigned to all treatments that are non-significantly different from (2), including (2) itself;

3. the iterated procedures stops when a rank is assigned to all treatments;

4. mean by columns (without considering zeros) provides a synthesis of the rank of each of the C treatments, it is a sort mid-rank;

5. finally to obtain the global ranking it is enough to apply the rank-transformation where, in the case of ties, the minimum value is

repeated (this because we used the convention that "the lower the rank, the better the treatment").

In order to better understand this procedure, we report here an example with $C = 8$ treatments:

Table 1. Example of a $C \times C$ matrix for the rank assignment rule.

Ord. tr.	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
(1)	1	1	1	0	0	0	0	0
(2)		1	1	1	1	0	0	0
(3)			1	1	1	1	0	0
(4)				1	1	1	1	0
(5)					1	1	1	1
(6)						1	1	1
(7)							1	1
(8)								1
Rank ass.								
1	1	1	1					
2		2	2	2	2			
3			3	3	3	3		
4				4	4	4	4	
5					5	5	5	5
Col.s mean	1	1.5	2	3	3.5	4	4.5	5
Rank	1	2	3	4	5	6	7	8

3.3. Resampling strategies

Here we describe the resampling approaches we have implemented to obtain the distribution of the ranking parameter estimators, they are based on either bootstrap or permutation resampling.

1. The first bootstrap version is suggested by the global null hypothesis: since under H_0^G the observations of all groups are exchangeable, resampling can be made (at every iteration) on the rows of the whole data-set, so the new dataset is considered as the actual

observed one. Hereafter we refer to this version as the *simple bootstrap*.

For the reason that we are interested in pairwise comparisons each comparison should only involve the observations belonging to the pair of treatments considered. This is because if we resample the whole data-set as in the simple version, observation from possibly inactive groups (so under H_0^G) could influence partial p-values, resulting in a lack of power of the tests (see Basso *et al.*, 2009).

With this objective we constructed a pseudo dataset before resampling: a 3-dimensional array on which layer j contains data from the j-th variable and every column of this layer contains the pooled vector of observations of the i-th and h-th treatment in the j-th variable (so there are K columns). Then rows of *this* pseudo dataset are resampled instead of the original one. Hereafter we refer to this strategy as the synchronized approach.

2. In the second bootstrap version we adopted the synchronized approach. The numerator of the statistic (8) on page 88 is calculated as in the same way of the previous version, while to obtain the $\widehat{\boldsymbol{\sigma}}_j^2$ on the denominator, the original structure of the dataset has to be recomposed at every iteration. So the first n values of columns of the type $\texttt{col}_0 = 1$, $\texttt{col}_i = \texttt{col}_{i-1} + (C - i)$, $i = 1, \ldots, C-2$ and the whole last column, must be pooled together before applying the ANOVA model for the $\hat{\boldsymbol{\sigma}}_j^2$ calculation. Hereafter we refer to this version as the *synchronized bootstrap*.

3. The first permutation strategy is similar to the precedent one: the same pseudo dataset is constructed but rows are resampled *without* replacement, so the reference distribution is the permutation one and therefore we are conditioning on the observed dataset. Nevertheless, in this version observed p-values are still calculated with reference to the $t_{C\cdot(n-1)}$ distribution. Hereafter we refer to this version as *synchronized permutation with parametric p-values*, because p-values are obtained as in the synchronized bootstrap case, but with reference to the permutation distribution.

4. The second permutation strategy has the same layout of the precedent but observed p-values used for the combining function are permutation p-values (see formula (10) on page 89). Hereafter we refer to this version simply as *synchronized permutation*.

We have decided to use the two permutation strategies only with the NPC method, since the synchronized approach is constructed specifically to increase the power of the permutation tests in a pairwise comparisons layout (see Basso *et al.* (2009), Pesarin and Salmaso (2010) and Pesarin (2001) for details on theory and motivations).

4. Simulation study

In this simulation study we analysed the behaviour of the proposed methods under the null and under the alternative hypothesis using these settings:

- generated data from the model: $y_{ijk} = \mu_{ij} + \varepsilon_{ijk}$ where μ_{ij} is the (known) mean of the i-th treatment in the j-th variable and ε_{ijk} are i.i.d random errors ($k = 1, \ldots, n$);
- 3 distributions for random errors: $\mathcal{N}(0, 1)$, $\mathcal{E}xp(1)$ and t_2 ;
- $C = 5$ number of treatments;
- $n = 8$ number of replications, i.e. sample size for every treatment;
- $p = 3,\ 6$ and 9 variables (only 3 variables under H_0^G);
- $MC = 1000$ datasets generated independently for each combination of the settings.

Hence there are 3 variable settings (one for each p) $\times$ 3 distributions for errors $\times$ 4 resampling strategies: the first 2 applied with either the NPC or the Mahalanobis method, the second 2 applied only with the NPC method so that we have 54 simulations to run under H_1^G and $6 \times 3 = 18$ simulations to run under H_0^G (just three for each method).

With the aim of checking the *type I error* probability of the proposed procedures, we perform a simulation for each procedure with 3 variables under H_0^G. Figure 1 reports the behavior of the observed rejection rate (based on the $MC = 1000$ replications) of the procedures.

While under H_0^G mean performances of the treatments are all equal, under H_1^G mean performances could differ in many ways. Therefore we have chosen to use an empirical rule for calibrating means, with the aim of keeping equal distances between groups when increasing dimensionality. This rule can be described as follows:

- let $\boldsymbol{\mu}_i = (\mu_{i1}, \ldots \mu_{ip})^\top$ and calculate $\boldsymbol{\mu}_i^\top \cdot \boldsymbol{\mu}_i$ for every "i", i.e. the squared Euclidean distance from the origin;

- try to align these distances (5 in our cases) on a straight line and calculate the slope, for example estimating the least square regression line of a linear model, hereafter we refer to this value as "$\boldsymbol{\eta}$";

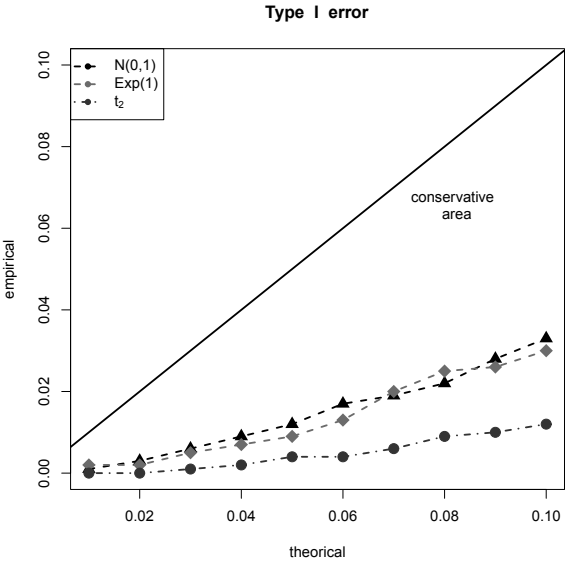
- go on modifying vectors of means until $\boldsymbol{\eta}$ is close enough to the chosen value for each of the 3 settings ($p = 3,\ 6$ and 9).

Hence $\boldsymbol{\eta}$ could be considered as a synthetic measure of the real distance between two consecutive groups (treatments). It is the core of the information that we wich to extract from the experiment, hence it has to be controlled: if it is too high (or low), the proportion of correct classified ranking could be too close to 1 (or 0), making impossible to effectively compare the proposed procedures.

With this aim we ran a simulation with 3 and 6 variables, with normal errors and decided to set all $\boldsymbol{\eta}$s as near as possible to the common value of -1.7. The following tables report the values of the true means calculated as previously described.

As a result of this calibration the global information on distances between groups remained the same in all simulations, hence increasing dimensionality do not increase global information and so, for a given variable, groups are closer in the simulations with 9 variables than in that with 3. This could be an explanation of the behavior of the exact ranking rates with the NPC method, in settings with the same distribution for errors and

***Mahalanobis** method,* simple *bootstrap.*



***Mahalanobis** method,* synchronised *bootstrap.*

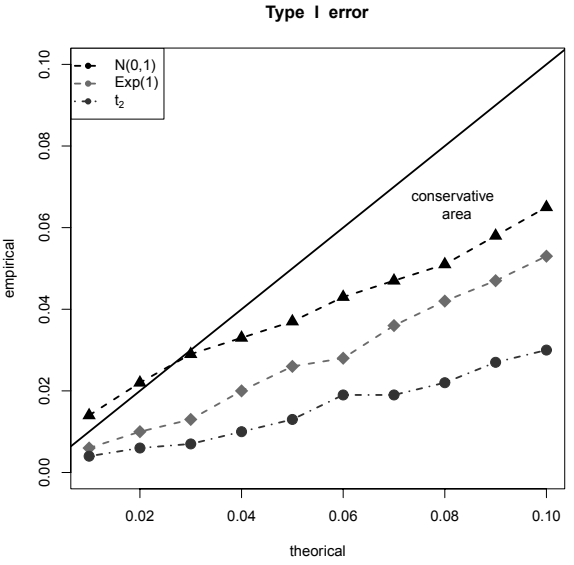
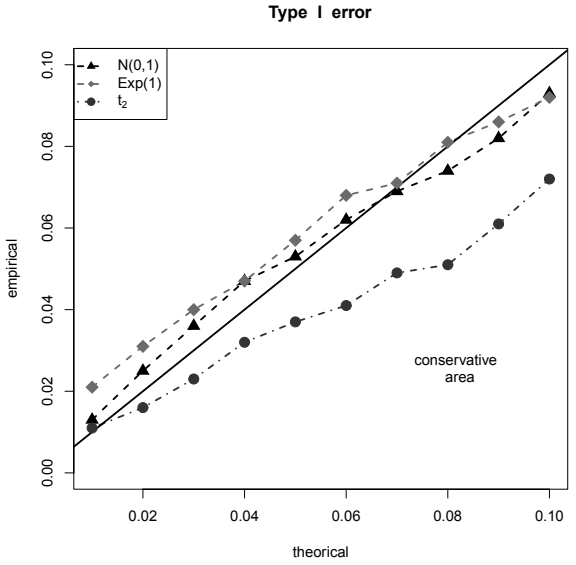


Figure 1. Behavior under H_0^G

***NPC** method,* simple *bootstrap.*



***NPC** method,* synchronised *bootstrap.*

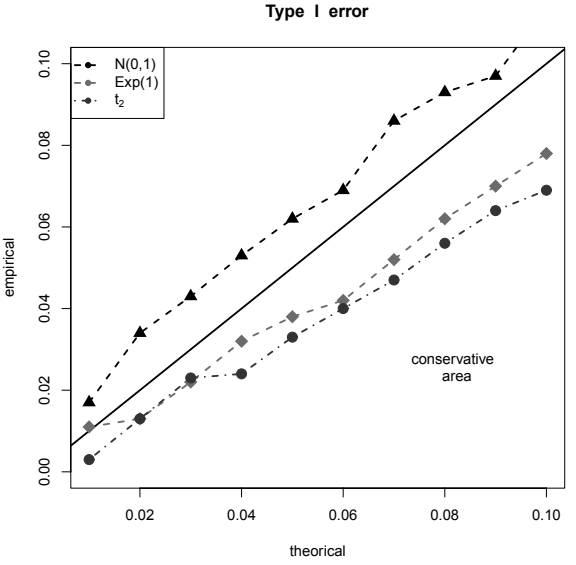
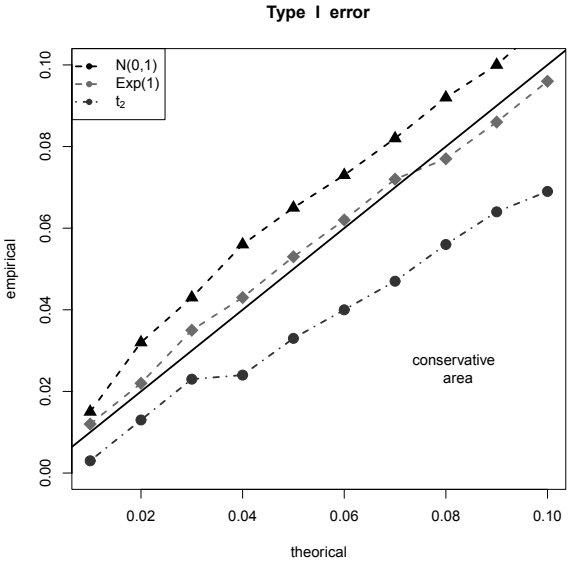


Figure 1. Behavior under H_0^G

***NPC** method,* synchronized *permutation with parametric p-values.*



***NPC** method and* synchronized *permutation.*

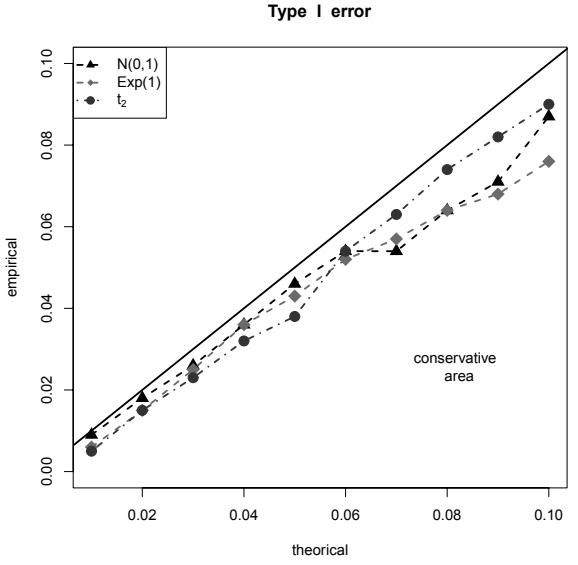


Figure 1. Behavior under H_0^G

same resampling strategy, when dimensionality increase. This does not seem to occur with the Mahalanobis method (see figure 2).

Table 2. True means under H_1

	$p = 3$ AND $\eta \cong -1.732$			
		$\boldsymbol{\mu}_1$	$\boldsymbol{\mu}_2$	$\boldsymbol{\mu}_3$
Group	1^{st}	90	89	88
	2^{nd}	88.5	88	87.5
	3^{rd}	87	87	87
	4^{th}	85.5	86	86.5
	5^{th}	84	85	86

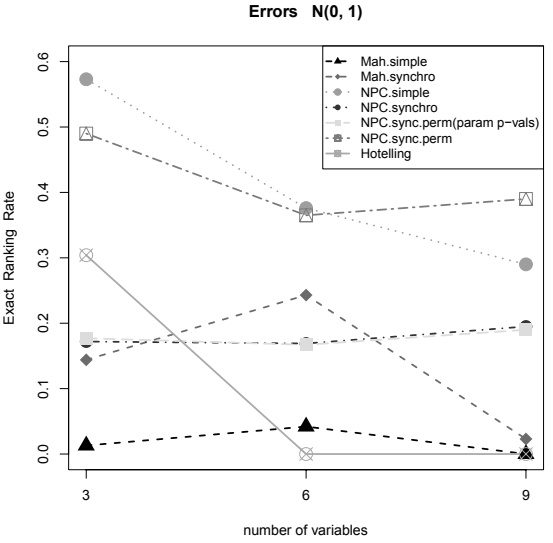
	$p = 6$ AND $\eta \cong -1.708$					
		$\boldsymbol{\mu}_1$	$\boldsymbol{\mu}_2$	$\boldsymbol{\mu}_3$	$\boldsymbol{\mu}_4$	$\boldsymbol{\mu}_5$	$\boldsymbol{\mu}_6$
Group	1^{st}	91	90	89	88	87	86
	2^{nd}	90.17	89	88.17	87.33	86.5	85.67
	3^{rd}	89.33	88	87.33	86.67	86	85.33
	4^{th}	88.5	87	86.5	86	85.5	85
	5^{th}	87.67	86	85.67	85.33	85	84.67

	$p = 9$ AND $\eta \cong -1.702$								
		$\boldsymbol{\mu}_1$	$\boldsymbol{\mu}_2$	$\boldsymbol{\mu}_3$	$\boldsymbol{\mu}_4$	$\boldsymbol{\mu}_5$	$\boldsymbol{\mu}_6$	$\boldsymbol{\mu}_7$	$\boldsymbol{\mu}_8$	$\boldsymbol{\mu}_9$
Group	1^{st}	91	90	90	89	89	88	88	86	86
	2^{nd}	90.78	89.56	89.33	88.11	88	87.22	87.44	85.67	85.78
	3^{rd}	90.56	89.11	88.67	87.22	87	86.44	86.89	85.33	85.56
	4^{th}	90.33	88.67	88	86.33	86	85.67	86.33	85	85.33
	5^{th}	90.11	88.22	87.33	85.44	85	84.89	85.78	84.67	85.11

In the next pages we report some tables resulting from the simulations study and, in order to summarise the main information of all tables (not given here), we also report figures of the *"exact ranking rate" vs* the number of variables (p) and of the number of correct ranking of the 3^{rd} level (the median level) for different type of errors (figure 2).

Starting from the figure related to the Mahalanobis method we can state that the use of the *synchronised approach* improve performances (with respect to the *simple approach*) in the sense that it increases the exact ranking rate.

Errors $\mathcal{N}(0,1)$,exact ranking rate.



Errors $\mathcal{N}(0,1)$ 3^{rd} level correct ranking.

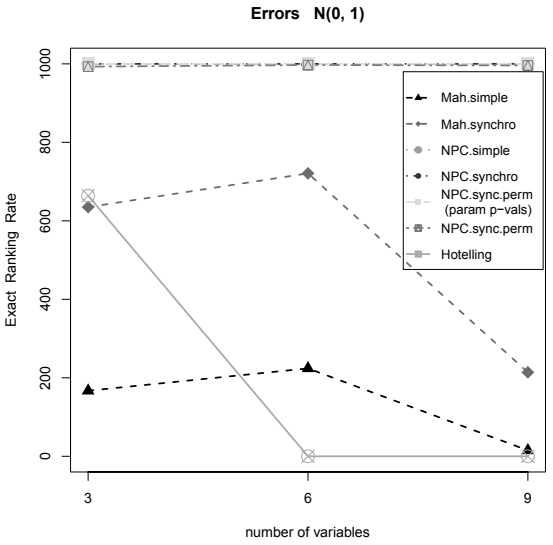
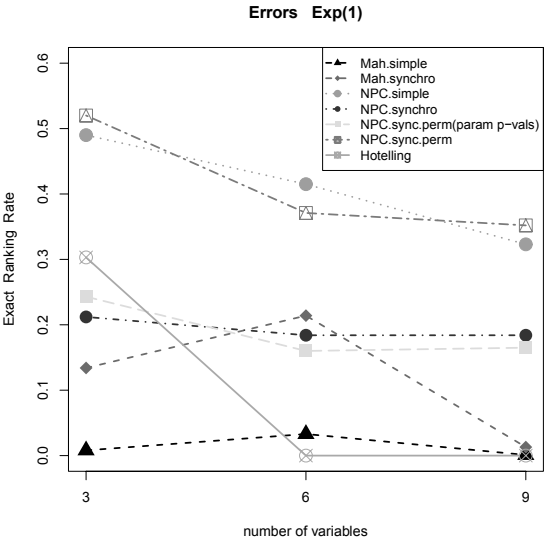


Figure 2. Exact ranking rate for the proposed procedures vs number of variables.

Errors $\mathcal{E}xp(1)$, exact ranking rate.



Errors $\mathcal{E}xp(1)$ 3^{rd} level correct ranking.

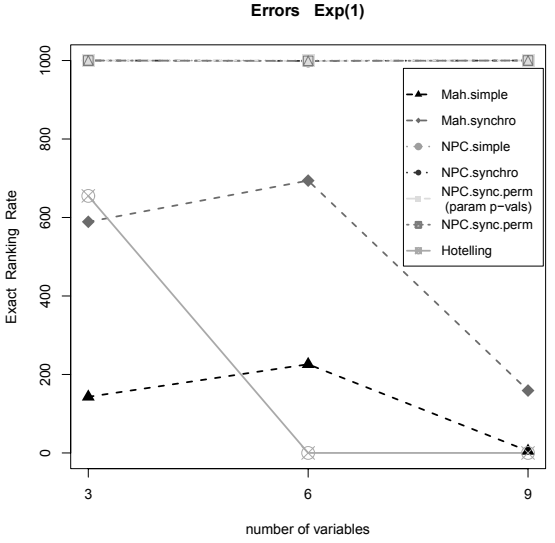
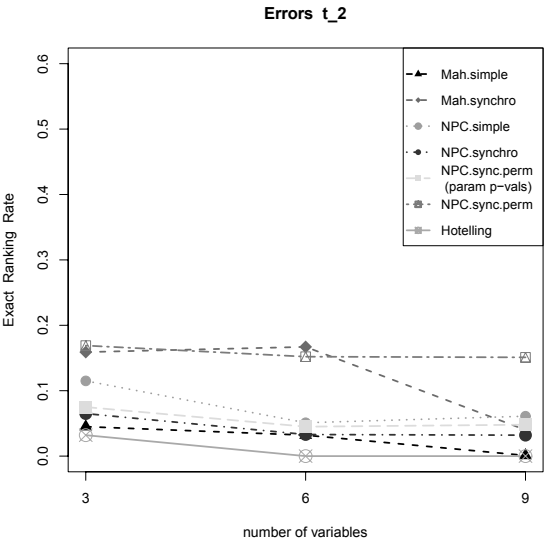


Figure 2. Exact ranking rate for the proposed procedures vs number of variables.

Errors $\mathbf{t}_2$*, exact ranking rate.*



Errors $\mathbf{t}_2$ 3^{rd} *level correct ranking.*

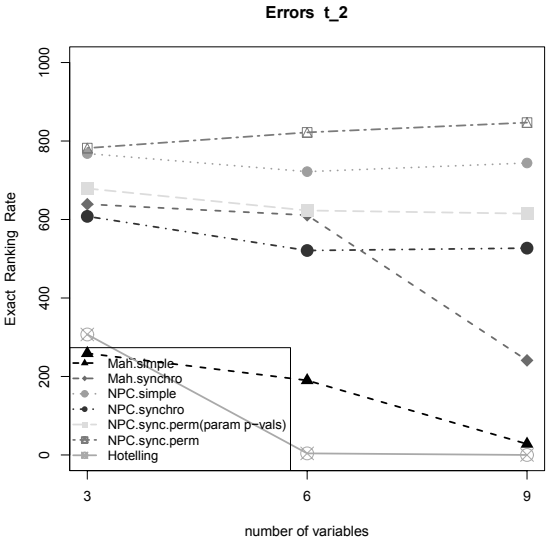


Figure 2. Exact ranking rate for the proposed procedures vs number of variables.

SYNCHRONIZED PERMUTATIONS (PARAMETRIC p-values)

Errors Type			P = 3					P = 6					P = 9				
			Estimated Rank														
			1st	2nd	3rd	4th	5th	1st	2nd	3rd	4th	5th	1st	2nd	3rd	4th	5th
N(0,1)	True Rank	1st	1000	0	0	0	0	1000	0	0	0	0	1000	0	0	0	0
		2nd	0	1000	0	0	0	0	1000	0	0	0	0	1000	0	0	0
		3rd	0	0	1000	0	0	0	0	1000	0	0	0	0	1000	0	0
		4th	0	0	10	986	4	0	0	64	925	11	0	0	81	900	19
		5th	0	0	10	813	177	0	0	64	769	167	0	0	81	729	190
		TOT.	1000	1000	1020	1799	181	1000	1000	1128	1694	178	1000	1000	1162	1629	209
	Exact Ranking Rate		0,177					0,167					0,190				
Exp(1)			1st	2nd	3rd	4th	5th	1st	2nd	3rd	4th	5th	1st	2nd	3rd	4th	5th
	True Rank	1st	1000	0	0	0	0	999	1	0	0	0	1000	0	0	0	0
		2nd	0	1000	0	0	0	1	999	0	0	0	0	1000	0	0	0
		3rd	0	0	1000	0	0	0	0	999	1	0	0	0	1000	0	0
		4th	0	0	5	986	9	0	0	48	931	21	0	0	39	931	30
		5th	0	0	5	752	243	0	0	47	792	161	0	0	39	796	165
		TOT.	1000	1000	1010	1738	252	1000	1000	1094	1724	182	1000	1000	1078	1727	195
	Exact Ranking Rate		0,243					0,160					0,165				
t_2			1st	2nd	3rd	4th	5th	1st	2nd	3rd	4th	5th	1st	2nd	3rd	4th	5th
	True Rank	1st	963	37	0	0	0	958	41	1	0	0	958	42	0	0	0
		2nd	148	809	42	0	1	172	791	32	4	1	157	806	34	3	0
		3rd	111	170	679	31	9	132	207	623	29	9	117	220	615	36	12
		4th	111	101	351	370	67	131	137	329	314	89	116	126	357	311	90
		5th	111	91	312	334	152	131	114	290	316	149	116	119	309	303	153
		TOT.	1444	1208	1384	735	229	1524	1290	1275	663	248	1464	1313	1315	653	255
	Exact Ranking Rate		0,075					0,045					0,048				

SYNCHRONIZED PERMUTATIONS (PERMUTATION p-values)

Errors Type			P = 3					P = 6					P = 9				
			Estimated Rank														
			1st	2nd	3rd	4th	5th	1st	2nd	3rd	4th	5th	1st	2nd	3rd	4th	5th
N(0,1)	True Rank	1st	996	4	0	0	0	998	2	0	0	0	998	2	0	0	0
		2nd	4	993	3	0	0	2	997	1	0	0	2	995	3	0	0
		3rd	0	4	993	3	0	0	2	997	1	0	0	3	996	1	0
		4th	0	1	45	953	1	0	0	64	933	3	0	0	74	918	8
		5th	0	1	42	459	498	0	0	62	571	367	0	0	73	536	391
		TOT.	1000	1003	1083	1415	499	1000	1001	1124	1505	370	1000	1000	1146	1455	399
	Exact Ranking Rate		0.490					0.365					0.390				
Exp(1)			1st	2nd	3rd	4th	5th	1st	2nd	3rd	4th	5th	1st	2nd	3rd	4th	5th
	True Rank	1st	999	1	0	0	0	1000	0	0	0	0	1000	0	0	0	0
		2nd	1	999	0	0	0	0	1000	0	0	0	0	1000	0	0	0
		3rd	0	0	1000	0	0	0	0	999	1	0	0	0	1000	0	0
		4th	0	0	18	981	1	0	0	20	975	5	0	0	20	974	6
		5th	0	0	16	462	522	0	0	19	609	372	0	0	20	628	352
		TOT.	1000	1000	1034	1443	523	1000	1000	1038	1585	377	1000	1000	1040	1602	358
	Exact Ranking Rate		0.520					0.371					0.352				
t_2			1st	2nd	3rd	4th	5th	1st	2nd	3rd	4th	5th	1st	2nd	3rd	4th	5th
	True Rank	1st	903	94	3	0	0	928	72	0	0	0	951	48	1	0	0
		2nd	116	821	62	1	0	82	873	45	0	0	53	905	42	0	0
		3rd	19	176	782	21	2	10	146	822	21	1	4	128	847	21	0
		4th	19	79	304	550	48	10	84	317	547	42	4	72	369	515	40
		5th	19	65	210	384	322	10	73	224	403	290	4	69	263	379	285
		TOT.	1076	1235	1361	956	372	1040	1248	1408	971	333	1016	1222	1522	915	325
	Exact Ranking Rate		0.169					0.152					0.151				

5. Discussions and conclusions

The aim of this work was developing a method capable of producing ranking of multivariate treatments, i.e. when considering the joint information from more than one response variable. We have compared several resampling strategies using two score-statistics in a simulation study that have highlighted some differences between the two proposed types of score, namely *NPC* and *Mahalanobis*, and provided some hints about differences between resampling approaches. The proposed nonparametric methods proved to be reliable tools to rank treatments from a multivariate point of view. With reference to the two proposed types of scores, we can state that the *NPC* method performs generally better than the other, especially in the *synchronised permutation* version and using permutation instead of parametric p-values. Under H_1 it has the highest exact ranking rate with all numbers of variables and all distributions for errors (even if there is a little crossing over with the 3 variables setting) as it can be seen comparing results in figure 2, whereas under H_0 it respects the α-nominal level better than the others (it is just below the nominal level, see figure 1) and seems to be not influenced by the type of error, in fact lines can be almost overlapped. This is probably due to the lack of power of partial tests in the simple resampling strategies as already pointed out. Moreover NPC method is quite flexible, in fact it has the advantage that the test statistics involved in its expression can be changed in accordance to the nature of the statistical problem at hand and therefore it naturally allows to deal with data from both (ordered or binary) categorical and continuous variable. NPC method is less affected by numerical problems than the Mahalanobis method, because it considers the problem variable-by-variable, hence it can be used even in situations where the number of variables exceeds the sample size of the experiment. In fact we can say that although it works with one variable at a time, it is able to effectively summarise information deriving from all variables. Morevoer, in the synchronised permutation approach the time of calculation is reduced 10–15 times using the permutation p-values in place of the parametric ones, and this could be useful in contexts where there is a low computational power such as multi-threaded web applications. As far as some directions for

future research are concerned, it could be useful to extend the proposed methodology to different designs (for example the randomised complete block design). Moreover, there is some further aspects that could be better investigate. First of all, the use of the synchronised approach improves the exact ranking rate with the Mahalanobis method but this does not occur with the NPC method with the bootstrap strategy. This is due to the 4^{th} and the 5^{th} treatments that the procedure consider as equal too many times. This aspect could be further explored by focusing on the effect of using the pairwise approach (a methodology developed inside the permutation context) with the bootstrap resampling strategy. Another interesting further development, especially in the field of evaluation of educational systems, could be the application (and then the evaluation) of the proposed procedures on datasets composed by (ordered or binary) categorical variables. Of course the statistics has to be changed in accordance to the nature of the specific variable (e.g. statistics of the goodness-of-fit type for ordered variables). Finally, additional research is needed to study the effect of heteroscedastic and/or dependent random errors and it could be worth to perform a simulation study where the main focus is on the sample size "n", with the aim of finding the minimum "n" such that the exact ranking rate is not lower than a fixed rate.

Acknowledgements: Authors wish to thank the University of Padova (CPDA 088513/08 and CPDA092350/09) and the Italian Ministry for University and Research (2008WKHJPK/002) for providing the financial support for this research.

References

AISE (2009). A.I.S.E. Detergent Test Protocol - 2009, www.aise.eu.

AISE (2010). A.I.S.E. web Site: www.aise.eu.

Arboretti G. R., Basso D., Bonnini S. and Corain L. (2008), A robust approach for treatment ranking within the multivariate one-way ANOVA layout, in P. Brito ed., *Proceedings in computational statistics*, 649–657.

Basso D., Pesarin F., Salmaso L. and Solari A. (2009), *Permutation tests for*

stochastic ordering and ANOVA: theory and applications with R, Springer, New York.

Bonnini S., Corain L., Cordellina A., Crestana A., Musci R. and Salmaso L. (2009), A novel global performance score with application to the evaluation of new detergents, in M. Bini, P. Monari, D. Piccolo, L. Salmaso (eds.), *Statistical methods for the evaluation of educational services and quality of products*, Physica-Verlag, Heidelberg, 161–179.

Bonnini S., Corain L. and Salmaso L. (2006), A new statistical procedure to support industrial research into new product development, *Quality and Reliability Engineering International*, 22, 555–566.

Box G. E. P. and Cox D. R. (1964), An analysis of transformations, *Journal of the Royal Statistical Society-Series B*, 26, 211–252.

Corain L. and Salmaso L. (2007), A nonparametric method for defining a global preference ranking of industrial products, *Journal of Applied Statistics*, 34, 203–216.

Draper, N. R. and Cox, D. R. (1969), On distributions and their transformation to Normality, *Journal of the Royal Statistical Society-Series B*, 31, 472–476.

Fayers P. M. and Hand D. J. (2002), Casual variables, indicator variables and measurement scales: an example from quality of life, *Journal of the Royal Statistical Association*, 165, 1–21.

Gupta Shanti S. and Panchapakesan S. (2002), *Multiple decision procedures*, 2nd ed., Siam, New York.

Hinkley D. V. (1975), On power transformations to symmetry, *Biometrika*, 62, 101–111.

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

Pesarin F. and Salmaso L. (2010), *Permutation tests for complex data*, Wiley & Sons, Chichester.

Westfall P. H., Tobias R. D., Rom D., Wolfinger R. D. and Hochberg Y. (1999), *Multiple comparisons and multiple tests using the SAS System*, NC: SAS System.