Journal of Statistical Software

January 2017, Volume 76, Issue 4. doi: 10.18637/jss.v076.i04

Nonparametric Inference for Multivariate Data: The R Package npmv

Amanda R. Ellis
University of Kentucky

Woodrow W. Burchett
University of Kentucky

Solomon W. Harrar
University of Kentucky

Arne C. Bathke
Universität Salzburg/University of Kentucky

Abstract

We introduce the R package **npmv** that performs nonparametric inference for the comparison of multivariate data samples and provides the results in easy-to-understand, but statistically correct, language. Unlike in classical multivariate analysis of variance, multivariate normality is not required for the data. In fact, the different response variables may even be measured on different scales (binary, ordinal, quantitative). p values are calculated for overall tests (permutation tests and F approximations), and, using multiple testing algorithms which control the familywise error rate, significant subsets of response variables and factor levels are identified. The package may be used for low- or high-dimensional data with small or with large sample sizes and many or few factor levels.

Keywords: MANOVA, multiple testing, closed testing procedure, rank test, permutation test, randomization test, familywise error rate.

1. Introduction

The paper introduces the R (R Core Team 2016) package **npmv** (Burchett and Ellis 2017) that provides valid inference procedures for samples of multivariate observations. The package is available from the Comprehensive R Archive Network (CRAN) at `http://CRAN.R-project.org/package=npmv`, and the underlying methodology is based on the nonparametric approach to multivariate inference presented in Bathke and Harrar (2008), Harrar and Bathke (2008a), Harrar and Bathke (2008b), Bathke, Harrar, and Madden (2008), Bathke, Harrar, and Ahmad (2009), and Liu, Bathke, and Harrar (2011). One major achievement in the recent methodology development is that no parametric assumptions such as multivariate normality have to

Sample 1				Sample 2				...	Sample a			
$X_{11}^{(1)}$	$X_{12}^{(1)}$	$\dots$	$X_{1n_1}^{(1)}$	$X_{21}^{(1)}$	$X_{22}^{(1)}$	$\dots$	$X_{2n_2}^{(1)}$	$\dots$	$X_{a1}^{(1)}$	$X_{a2}^{(1)}$	$\dots$	$X_{a,n_a}^{(1)}$
$X_{11}^{(2)}$	$X_{12}^{(2)}$	$\dots$	$X_{1n_1}^{(2)}$	$X_{21}^{(2)}$	$X_{22}^{(2)}$	$\dots$	$X_{2n_2}^{(2)}$	$\dots$	$X_{a1}^{(2)}$	$X_{a2}^{(2)}$	$\dots$	$X_{a,n_a}^{(2)}$
	$\dots$				$\dots$			$\dots$		$\dots$		
$X_{11}^{(p)}$	$X_{12}^{(p)}$	$\dots$	$X_{1n_1}^{(p)}$	$X_{21}^{(p)}$	$X_{22}^{(p)}$	$\dots$	$X_{2n_2}^{(p)}$	$\dots$	$X_{a1}^{(p)}$	$X_{a2}^{(p)}$	$\dots$	$X_{a,n_a}^{(p)}$

Table 1: General schematic layout for multivariate observations from several samples.

be made. Such assumptions, which are required for the classical parametric MANOVA (multivariate analysis of variance), itself available through the standard R package **stats** (`manova` function), are rather restrictive and hard to verify. In fact, they are arguably one of the main reasons why classical MANOVA is rarely used: It is almost impossible to justify its prerequisites. Another limitation of the classical MANOVA is that even when the assumptions of multivariate normality are met, MANOVA tests typically provide answers that are not useful in practice: They only make a global statement about significance. Classical MANOVA procedures do not provide coherent information about which sub-groups of response variables or factor levels are responsible for the global significance. The R package **npmv** solves both problems by (a) providing a fully nonparametric approach and (b) supplementing the global test with a comprehensive procedure identifying significant response variables and factor levels – while at the same time controlling the familywise error rate.

1.1. Nonparametric multivariate model

Consider the situation involving a samples (factor levels) of p-variate observation vectors (i.e., p response variables), with individual sample sizes $n_1, \dots, n_a$, respectively, and total sample size $N = \sum_{i=1}^{a} n_i$. The nonparametric model underlying the R package **npmv** simply states that the multivariate observation vectors $\mathbf{X}_{ij} = (X_{ij}^{(1)}, \dots, X_{ij}^{(p)})^\top$ are independent, and within the same factor level i, they follow the same p-variate distribution: $\mathbf{X}_{ij} \sim F_i$. Here and in the following, the different variables are denoted by $k = 1, \dots, p$, the different conditions (treatments, sub-populations, factor levels) are indexed by $i = 1, \dots, a$, and within each condition, the n_i subjects (experimental units), on which the p-variate observations are made, are indexed by $j = 1, \dots, n_i$, We assume implicitly, that the same p response variables are observed at each of the a factor levels, and these p variables may be dependent. The dependence structure does not have to be specified. Also, the marginal distributions may of course be different for the different response variables.

Typical global statistical hypotheses in this context are the following. "Are the a samples from the same population (multivariate distribution)?" or "Do the a treatments have the same effect?" These can be formulated as $H_0 : F_1 = \ldots = F_a$. Further, if a global hypothesis is rejected, investigators would like to know *which* variables or treatments contributed to the significant overall effect. The R package **npmv** provides a comprehensive way to answer these questions and to summarize the results.

Table 1 shows a schematic layout for the type of data considered in the package. Different rows correspond to the p different variables, different treatment groups are indicated by the blocks titled *Sample i*, and each column represents one experimental unit (subject, person).

For the special case of univariate responses (or, e.g., variable-wise inference for the different

Treatment	Replication	Weight	Botrytis	Fungi	Rating
3	1	6.90	4.0956	17.2355	1.0
3	2	8.30	5.1348	5.6482	1.0
3	3	8.40	6.0698	8.8012	1.5
3	4	7.95	2.7174	9.5109	1.5
6	1	8.60	1.1945	17.0649	1.0
6	2	8.50	0.5533	12.8631	1.0
6	3	8.20	0.7353	6.7647	0.5
6	4	9.50	0.9929	1.8440	1.0
8	1	6.20	4.2857	4.6428	1.0
8	2	9.00	1.5640	3.0303	3.0
8	3	6.80	0.8757	5.6042	0.0
8	4	8.50	2.4249	8.6605	2.0
9	1	7.50	15.5975	13.0817	1.0
9	2	6.70	10.2819	14.4279	1.0
9	3	8.70	13.2895	10.9211	2.5
9	4	7.40	18.3824	16.0295	3.0

Table 2: Strawberry data.

response variables), modern nonparametric inference methods have been implemented recently in the R package **nparcomp** (Konietschke, Placzek, Schaarschmidt, and Hothorn 2015). Also, an R implementation exists for a nonparametric analysis in the special case of repeated measurements, i.e., for the case where the p different variables constitute repeated observations of the *same* characteristic on the same subject, and are measured in the same units. In such a situation, the R package **nparLD** (Noguchi, Gel, Brunner, and Konietschke 2012) can be used. However, these tools available in the packages **nparcomp** and **nparLD** are not in general applicable to multivariate data which usually involve different, typically dependent characteristics measured in rather different units.

1.2. Examples

Strawberry data

The strawberry data set is a multivariate response data set that gives the measurements of weight, the percentage of Botrytis, percentage of other fungal species, and a rating of symptoms from Phomopsis leaf blight, for four plots of strawberries each treated with one of four treatments. Three of the treatments were different fungicides, and one was a control. The full data is listed in Table 2. Detailed descriptions of the data set are provided by Horst, Locke, Krause, McMahon, Madden, and Hoitink (2005) and Bathke *et al.* (2008).

The R package **npmv** includes the dataframe `sberry`, which provides the data from Table 2. Researchers were interested in finding out whether there was a difference between treatments, and, if so, on which response variables and particularly between which treatments. Note that the data contains one ordinal variable (rating) and three quantitative variables.

Anderson and Fisher's iris data

This classical data set contains the measurements of four quantitative variables, sepal length and width, as well as petal length and width, respectively, for 50 flowers from each of the three iris species *Iris setosa*, *Iris versicolor*, and *Iris virginica*. It is referred to as "Fisher's iris data" or "Anderson's iris data" (Fisher 1936; Anderson 1935), and available in the R package **datasets** as dataframe `iris`. Thus, these well-known data are easily accessible to every R user, and as a naturally multivariate data set, they provide a convenient and fitting example for the application of the R package **npmv**. The data set has become famous as an example for discriminant analysis (including the case where the species allocation of the observations is not given and needs to be estimated), thus natural questions to be answered in the context of multivariate inference are whether the different species can at all be differentiated through the four variables, which of the species differ from each other, and with respect to which of the variables.

1.3. Global multivariate test statistics

In order to test the overall null hypothesis that the multivariate distributions F_i, $i = 1, \ldots, a$, do not differ across the factor levels, test statistics using sums of squares and cross-products based on ranks are employed. Here, the ranks are taken variable-wise. As a consequence, the resulting test statistics are invariant under strictly monotone transformations of individual response variables. This is an important and desirable property, as, for example, changing the scale of one variable from percentage to proportion or from metric to imperial units, or using a different number set for an ordinal characteristic, should not change the results of the test. Details regarding the underlying theory can be found in Bathke *et al.* (2008) and Liu *et al.* (2011), and the references cited therein. Roughly speaking, in the underlying theoretical articles, rank-based analogs to classical multivariate tests have been defined, their asymptotic distributions derived, small sample approximations developed, and the comparative performance of different approximations has been investigated by means of extensive simulation studies. In Appendix A, we briefly summarize how the four rank-based test statistics of ANOVA type, Wilks' Lambda type, Lawley Hotelling type, and Bartlett Nanda Pillai type are constructed.

In addition to the F-distribution approximations that are provided (see Appendix A), each of the four test statistics is also used as the basis for a multivariate permutation or randomization test. To this end, the N data vectors are permuted, and the multivariate test statistics recalculated each time. For each of the four tests, these resulting values form the respective distribution whose quantiles are used to determine the p value of the corresponding permutation test (if all $N!$ permutations are performed) or randomization test (if a predetermined number of random permutations is performed). For the latter, the user can specify the number of permutations in the R package. Default is 10,000 permutations. The four test statistics mentioned above are also used in the subset algorithm explained in the next section.

Alternative methods for inference on multivariate data are available, for example, through the function `manova` in the standard R package **stats**. This function calculates classical normal theory MANOVA test statistics and the corresponding p values. It relies on the assumptions of equal covariance matrices for the different groups, and multivariate normality, and due to these restrictions, its use is very limited. Permutation and randomization tests are also implemented in other R packages, such as **coin** (Hothorn, Hornik, Van de Wiel, and Zeileis

2008), **energy** (Rizzo and Szekely 2016), **lmPerm** (Wheeler 2016), and **vegan** (Oksanen *et al.* 2016).

The first two packages can also be used to calculate global permutation test statistics for multivariate data (see also Hothorn, Hornik, Van de Wiel, and Zeileis 2006; Székely and Rizzo 2004). In comparison, **npmv** provides more detailed information than just the result from a global test: It also includes an algorithm to detect the sub-groups of response variables or factor levels that are responsible for the global significance.

Published applications of the package **npmv** can be found, for example, in Nardone *et al.* (2014, 2015), and Grabcanovic-Musija *et al.* (2015).

1.4. Which test to use?

Altogether, there are eight tests (four types, each with F approximation and as permutation test). None of these is uniformly better than the others. On the bright side, all of them will also typically be in good agreement. However, there will certainly be situations where the results differ slightly. We are providing the following advice, based on several simulation studies (cf. also the articles mentioned at the beginning of this section). This recommendation is the default setting in the R package **npmv**. See Section 3 on how to change those default settings.

1. For all situations where it can be used, Wilks' Lambda is used.
2. For high-dimensional data, the only test that can always be used, is the ANOVA-type statistic. Therefore, it is used whenever Wilks' Lambda is not available.
3. Currently, for $N < 10$, the permutation test is performed. For $10 \leq N < 30$, the randomization test is performed with 10,000 randomly chosen permutations. For $N \geq 30$, the F approximation is used.

2. Subset algorithm

After a rejection of the global hypothesis, researchers typically ask the following questions.

1. Which of the p variables displayed significant differences?
2. Which of the a factor levels contributed to the significant result?

In order to answer those questions, package **npmv** performs an all-subsets algorithm regarding variables and regarding factor levels, whenever computing time allows (i.e., whenever p and a are not too large).

2.1. Illustration of the procedure

The algorithm maintains control of the familywise error rate (default in the R package **npmv**: $\alpha = 0.05$). To this end, for factor level comparisons, the closed multiple testing principle (Marcus, Peritz, and Gabriel 1976; Sonnemann 2008) is used. For comparisons using different sets of variables, the closed multiple testing principle cannot be applied, and thus the multiple testing procedure adjusts the α-levels appropriately to ensure that strong control of the familywise error rate is guaranteed.

As an illustration of the closed testing procedure for factor levels, consider an example with four treatments. The closed multiple testing principle demands that the hypothesis stating equality of treatments 1 and 2, $H_0^{(1,2)} : F_1 = F_2$, may only be rejected if, in addition to this hypothesis, all other subset hypotheses are rejected that involve at least these two groups. That is, in addition to $H_0^{(1,2)}$, also the following hypotheses have to be rejected: $H_0^{(1,2,3)} : F_1 = F_2 = F_3$, $H_0^{(1,2,4)} : F_1 = F_2 = F_4$, and $H_0^{(1,2,3,4)} : F_1 = F_2 = F_3 = F_4$, as well as $H_0^{(1,2)(3,4)} : (F_1 = F_2) \wedge (F_3 = F_4)$.

Based on this principle, it is clear that the algorithm starts with the global multivariate test (all p variables, all a factor levels) at level α. Default setting for the algorithm is to use the ANOVA-type test with F approximation for each subset, but the user can choose any one from the eight available tests. Note that once this choice is made, the same test will be used for all subset tests.

If the global test rejects, the two-stage subset procedure starts. Otherwise, no further testing will be performed. By default, if $p \leq a$, the algorithm uses subsets of the variables, otherwise subsets of the factor levels. The user can specify to perform both. Assuming $p \leq a$, the multivariate test is now performed for all subsets with $p - 1$ variables, next for those subsets with $p - 2$ variables that satisfy further testing under the principle of logical coherence, and so forth. At the end of this stage, the user obtains the output of the procedure in the least redundant form. In case of testing subsets of the a factor levels, the algorithm stops after calculating the test for all pairs, as it does not make sense to consider single factor levels in order to formulate a meaningful hypothesis. However, it does make sense to consider single variables. The latter simply corresponds to a univariate analysis.

The last hypothesis mentioned in the illustration above is one of the *partition hypotheses*, which need to be accounted for when testing subsets of factor levels *and* when $a > 3$. These hypotheses are typically not tested explicitly because their number, which can be calculated using the Stirling numbers of the second kind, grows at a much faster rate than that of all other hypotheses combined. Indeed, for $a = 10$, the number of partition hypotheses is greater than 1.1×10^5, and for $a = 13$, it is already greater than 2.7×10^7. We have implemented a Bonferronization method sometimes referred to as *Ryan adjustment* to account for the partition hypotheses, while still maintaining strong control of the familywise error rate (see, e.g., Hommel 1985). Its use is illustrated in the following example.

Example

Assume that $a = 4$, $p = 6$. The notation $H_0^{i_1,\ldots,i_k}$ stands in the following for the null hypothesis involving the factor levels $i_1, \ldots, i_k$ and all $p = 6$ variables. Consider the following situation.

- $H_0^{(1,2,3,4)}$ rejected at level α.
- $H_0^{(1,2,3)}$, $H_0^{(1,2,4)}$, $H_0^{(1,3,4)}$ rejected at level $\frac{3}{4}\alpha$, $H_0^{(2,3,4)}$ not rejected at level $\frac{3}{4}\alpha$.

  The factor $\frac{3}{4}$ in this step reflects the adjustment that is necessary when testing subsets of factor levels for $a \geq 4$, and it stems from the fact that in this step, sets of three samples are considered, while the total consists of four samples.

- $H_0^{(1,2)}$, $H_0^{(1,3)}$ rejected at level $\frac{2}{4}\alpha$,
- $H_0^{(1,4)}$ not rejected at level $\frac{2}{4}\alpha$,
- $H_0^{(2,3)}$, $H_0^{(2,4)}$, $H_0^{(3,4)}$ not tested due to closed multiple testing principle.

The output being returned to the user by **npmv** states the significant subsets of factor levels as $\{1,2\}$, $\{1,3\}$, $\{1,2,3\}$, $\{1,2,4\}$, $\{1,3,4\}$, $\{1,2,3,4\}$. Due to the logic of the procedure, this summary contains the complete information about all individual tests that have (and have not) been performed. However, the package developers find it important to phrase the output in such a way that it can be understood and used by a wider range of researchers with basic statistical understanding. Thus, the results are not rendered as formulas, but instead using statistically correct standard language that can even be inserted directly into research papers. For details, see the examples in Section 3.

The maximum number of tests that could be performed in this procedure is $2^{\min(a,p)} - 1$. This should be kept in mind when using the procedure in case of large number of factor levels a *and* variables p.

3. How to use the R package npmv

The package **npmv** provides the R functions `nonpartest` and `ssnonpartest`, both used to compute nonparametric test statistics. The function `nonpartest` computes the global nonparametric test statistics, their permutation (randomization) test analogs, and calculates nonparametric relative effects, in addition to providing appropriate data visualization. The function `ssnonpartest` identifies significant subsets of variables and factor levels using a multiple testing algorithm that controls the familywise error rate. Below, we discuss the two functions, `nonpartest` and `ssnonpartest`, separately.

3.1. `nonpartest` function

Function `nonpartest` is used to perform global inference for several multivariate samples, along with providing appropriate descriptive statistics. In order to analyze the strawberry data described in Section 1.2 (and included with the package **npmv**), the following R code can be executed, after installing and loading the package **npmv**, which will import package **Formula** (Zeileis and Croissant 2010) that is being used in **npmv**.

```
R> data("sberry", package = "npmv")
R> nonpartest(weight | bot | fungi | rating ~ treatment, data = sberry,
+    permreps = 1000)
```

Note that **npmv** facilitates model equation input using the class '`formula`', with multiple response variables, followed after the tilde (`~`) by a single explanatory variable. The response variables may be metric, ordinal, or even binary. The following is a generic function call with default arguments, followed by an explanation of each of the arguments of the function.

```
R> nonpartest(formula, data, permtest = TRUE, permreps = 10000,
+    plots = TRUE, tests = c(1, 1, 1, 1), releffects = TRUE)
```

1. `formula` is an object of class '`formula`', with a single explanatory variable and multiple response variables (or an object that can be coerced to that class).
2. `data` is an object of class '`data.frame`', containing the variables in the formula.
3. `permtest` controls whether p values for the permutation (randomization) test are returned (default `TRUE`).
4. `permreps` specifies the number of replications in the randomization version of the permutation test (default 10,000).
5. `plots` allows the user to decide whether plots of the data shall be produced. If `TRUE`, then box-plots ($\max_i n_i > 10$) or dot-plots ($\max_i n_i \leq 10$) are generated, separately for each response variable (default `TRUE`). Standard graphical parameters to be passed on to the plot can simply be added as function arguments. As an example, adding `, col = "blue", las = 2` after `releffects = TRUE` specifies the plot color to be blue, and labels to be perpendicular to the axes.
6. `tests` is a vector of zeros and ones specifying which test statistics are to be calculated. A one corresponds to the respective test statistics to be returned. Default is for all four types of test statistics to be computed. The entries of this vector refer to the types in the following order: ANOVA type, Lawley-Hotelling type, Bartlett-Nanda-Pillai type, and Wilks' Lambda type.
7. `releffects` controls whether nonparametric relative treatments effects (see below) are to be calculated as an appropriate numerical descriptive measure complementing the inferential analysis (default `TRUE`).

In a first step, the function `nonpartest` checks to ensure that the data set does not have missing values, since the current state of the art of the nonparametric methods implemented in the code requires complete data with no missing values. Then, data visualizations are produced by default (they can be turned off, see above), in order for the user to have a visual comparison between the treatments for the different response variables. As an example, Figure 1 displays the dot-plot for the variable "Botrytis" in the strawberry data set. The package developers consider visual inspection of the data a high priority for every data analysis, and thus the function `nonpartest` provides the plots by default *before* any numerical descriptive and inferential results are shown.

Next, the function computes the chosen nonparametric test statistics and returns a list of test statistic values, as well as numerator and denominator degrees of freedom, and the p values for each statistic using both F approximation and permutation (randomization) method. The output for the strawberry data is as follows. It should be noted that the package output displays "Permutation Test p-value" instead of "P.T. p-value" in the last column, as well as "McKeon approx. for the Lawley Hotelling Test" and "Muller approx. for the Bartlett-Nanda-Pillai Test" instead of "LH Test" and "BNP Test", respectively. We have shortened these here to fit the output onto the page.

```
$results
                Test Statistic        df1     df2   P-value  P.T. p-value
ANOVA type test          2.984      6.836  27.343     0.019         0.006
```

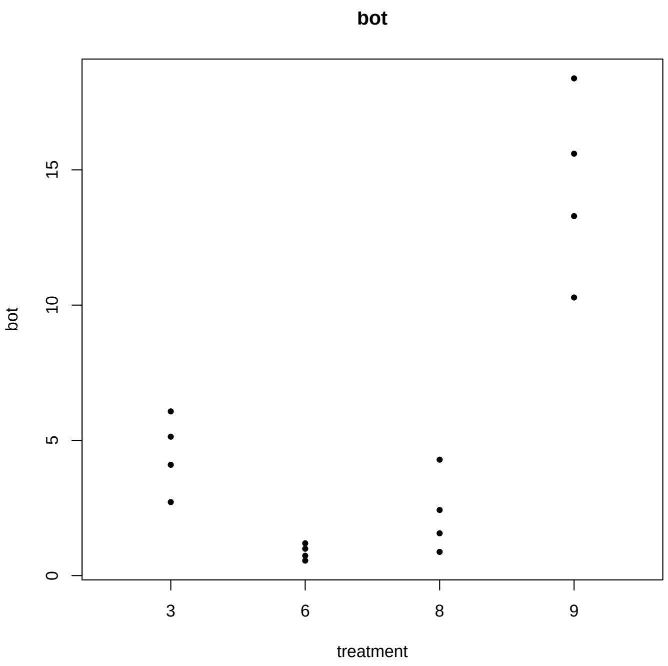

Figure 1: Dot-plot of "Botrytis" (vertical) vs. treatment (horizontal) in strawberry data.

```
LH Test                     5.769     12.000  12.000     0.002          0.001
BNP Test                    2.501     15.967  41.164     0.009          0.005
Wilks Lambda                4.166     12.000  24.103     0.001          0.002
```

Clearly, in this example, the treatment effect is highly significant, and there is good agreement between all eight tests that are provided with the package **npmv**. Finally, as a numerical description fitting the nonparametric paradigm, the empirical nonparametric relative treatment effects are listed for each variable.

```
$releffects
   weight     bot   fungi  rating
3  0.4375  0.5938  0.5625  0.5313
6  0.7266  0.1563  0.4843  0.3047
8  0.4453  0.3750  0.2188  0.5313
9  0.3906  0.8750  0.7344  0.6328
```

The relative effects quantify the tendencies observed in the data in terms of probabilities. For example, the plants in treatment group 9 tend to larger values compared to other treatment groups. The probability that a randomly chosen plant from group 9 exhibits a larger percentage of Botrytis than a randomly chosen plant from the full trial (including group 9) is 0.875. This is the maximum possible relative effect for this configuration (see, e.g., Brunner, Domhof, and Langer 2002; Acion, Peterson, Temple, and Arndt 2006, or Kieser, Friede, and Gondan 2013 for a detailed explanation of nonparametric relative treatment effects and their interpretation), which is in accordance with the display in Figure 1. Generally, the relative treatment effect (RTE) of treatment "k" is defined as the probability that a randomly chosen subject from treatment "k" displays a higher response than a subject that is randomly chosen from any of the treatment groups, including treatment "k".

Anderson and Fisher's iris data described also in Section 1.2 is available in the R package **datasets** as dataframe `iris`. Global nonparametric multivariate inference can be carried out using the following line of code, which selects all four response variables in the data set as dependent variables for the analysis, while the factor "Species" is selected as explanatory variable.

```
R> data("iris", package = "datasets")
R> nonpartest(Sepal.Length | Sepal.Width | Petal.Length | Petal.Width ~
+    Species, data = iris, permreps = 1000)
```

For a larger data set, such as `iris` (50 observations in each of the three groups), the function `nonpartest` automatically chooses box-plots in lieu of dot-plots to display the data (for brevity not rendered in this manuscript). In a next step, the inferential results are provided.

```
$results
                Test Statistic     df1       df2  P-value   P.T. p-value
ANOVA type test        178.511   3.826   281.234        0              0
LH Test                316.457   8.000   203.402        0              0
BNP Test                67.965   8.162   293.891        0              0
Wilks Lambda           155.763   8.000   288.000        0              0
```

Clearly, the difference between the three multivariate distributions is highly significant, according to each of the four test criteria. The relative effects provided in the next part of the output show that the differences between the three species are indeed quite pronounced, with regard to each variable. For example, the probability that a randomly chosen measurement of "Petal length" or "Petal width" from species *Iris setosa* is larger than a randomly chosen observation from the full trial (including *Iris setosa*) is 1/6, which is the minimum possible effect for this configuration. In general, the minimum and maximum possible effects for group i are $n_i/(2N)$ and $1-n_i/(2N)$, respectively. Thus, in this case, the minimum and maximum values are $50/300 = 1/6$ and $5/6$, respectively. In other words, each of the variables "Petal length" and "Petal width" perfectly discriminates *Iris setosa* from the other two species. Those two in turn also exhibit rather strong differences between them, so that the extreme inferential results become quite comprehensible.

```
$releffects
           Sepal.Length  Sepal.Width  Petal.Length  Petal.Width
setosa          0.19427      0.75307       0.16667      0.16667
versicolor      0.54767      0.30387       0.50593      0.50653
virginica       0.75807      0.44307       0.82740      0.82680
```

3.2. `ssnonpartest` function

The function `ssnonpartest` provides a more detailed comparison of the different multivariate samples using a subset algorithm that determines which of the variables or factor levels, respectively, contribute to the significant differences. For the strawberry data example, the function can be evoked using the following code.

```
R> data("sberry", package = "npmv")
R> ssnonpartest(weight | bot | fungi | rating ~ treatment, data = sberry,
+    test = c(1, 0, 0, 0), alpha = 0.05, factors.and.variables = TRUE)
```

Again, model equation input is based on the class 'formula'. Multiple response variables can be entered, followed by one single explanatory variable. Response variables may again be metric, ordinal, or even binary, as in the function nonpartest. A generic call with all arguments and, where applicable, their defaults, looks as follows.

```
R> ssnonpartest(formula, data, test = c(1, 0, 0, 0), alpha = 0.05,
+    factors.and.variables = FALSE)
```

1. formula is an object of class 'formula', with a single explanatory variable and multiple response variables (or an object that can be coerced to that class).

2. data is an object of class 'data.frame', containing the variables in the formula.

3. test is a vector of zeros and exactly one one specifying which test statistic is to be calculated for each of the subset tests. A one corresponds to the respective test statistic to be used throughout the subset testing procedure. The order of the test statistics is: ANOVA type, Lawley Hotelling type (McKeon's F approximation), Bartlett-Nanda-Pillai type (Muller's F approximation), and Wilks' Lambda type. Default is for the F approximation of Wilks' Lambda to be calculated wherever possible.

4. alpha (numerical) is the familywise level of significance at which hypothesis tests are to be performed (default 0.05).

5. If factors.and.variables is TRUE, then the subset algorithm is run both by factor levels and by variables (default FALSE).

In the same way as the nonpartest function, the ssnonpartest function also checks to ensure that there are no missing data. The function outputs all those subsets that turned out significant. For the strawberry data, the following output is produced by the function ssnonpartest.

```
The ANOVA type statistic will be used in the following test
The Global Hypothesis is significant, subset algorithm will continue

~Performing the Subset Algorithm based on Factor levels~
The Hypothesis of equality between factor levels  3 6 8 9 is rejected
The Hypothesis of equality between factor levels  6 8 9 is rejected
The Hypothesis of equality between factor levels  3 6 9 is rejected
All appropriate subsets using factor levels have been checked using a closed
  multiple testing procedure, which controls the maximum overall type I
  error rate at alpha= 0.05

~Performing the Subset Algorithm based on Response Variables~
The Hypothesis of equality using response variables
```

```
  weight bot fungi rating is rejected
The Hypothesis of equality using response variables
  bot fungi rating is rejected
The Hypothesis of equality using response variables
  weight bot fungi is rejected
The Hypothesis of equality using response variables  bot fungi is rejected
All appropriate subsets using response variables have been checked using a
  multiple testing procedure, which controls the maximum overall type I
  error rate at alpha= 0.05
```

Here, the first line of the output indicates that the global hypothesis test has shown a significant result using *all* variables and *all* factor levels. This test serves as a "gatekeeper" – only when it is significant, any further subset testing will be done. Consequently, if the global hypothesis had not been significant, the output would have indicated so, and the subsets would not have been tested.

By our specification in the function call, for this global test, as well as for all the subset tests, Wilks' Lambda is chosen as the test statistic. Only one of the four available test types can be chosen for the procedure, and once chosen, this type is used for all subsets to keep the results comparable. In order to keep runtime low, the F approximation is used to approximate the sampling distribution, rather than the computationally more intensive permutation (randomization) method. The default for the function is to only consider subsets comprised of either factor levels or variables based on the conditions previously mentioned (basically whichever leads to the smaller number of subset tests).

For illustrative purposes, and since for this data example it is resonable to do so, we have chosen the option to perform subset testing for both types of subsets in the strawberry data. Here, the first set of tests is based on factor levels. Only the significant subsets are listed, and for factor levels the smallest subsets considered are those of size two. In this example, a significant result is obtained between every set of three factor levels, as well as between the two treatments "3" and "6"'. Thus, the procedure has provided a comprehensive list of significances between treatments, while maintaining the familywise error rate at the (default) α-level of 5%. Namely, only the difference between factor levels "3" and "6"' is significant. The last portion of the output shows the results of testing the subsets based on response variables. Similar to the factor levels only the signifanct subsets are listed. However, for response variables it makes sense to look at subsets comprised of only one variable. In this example, the variable "Botrytis" turns out significant all by itself, as well as in combination with every other variable. Among the four response variables considered, only "Botrytis" has been shown to exhibit results differing significantly between the treatments. Note that the wording provided in the output can be understood immediately by researchers with basic statistical knowlegde, and it could even be inserted verbatim into a research paper.

Now considering the iris data example, the descriptive output provided by the `nonpartest` function above suggests that there are marked differences between all three species, visible in all four variables. We investigate this using the subset testing procedure which can be evoked with the following code.

```
R> data("iris", package = "datasets")
R> ssnonpartest(Sepal.Length | Sepal.Width | Petal.Length | Petal.Width ~
+    Species, data = iris, test = c(1, 0, 0, 0), alpha = 0.05,
+    factors.and.variables = TRUE)
```

It provides the following output.

```
The ANOVA type statistic will be used in the following test
The Global Hypothesis is significant, subset algorithm will continue

~Performing the Subset Algorithm based on Factor levels~
The Hypothesis of equality between factor levels
  setosa versicolor virginica is rejected
The Hypothesis of equality between factor levels
  versicolor virginica is rejected
The Hypothesis of equality between factor levels
  setosa virginica is rejected
The Hypothesis of equality between factor levels
  setosa versicolor is rejected
All appropriate subsets using factor levels have been checked using a closed
  multiple testing procedure, which controls the maximum overall type I
  error rate at alpha= 0.05

~Performing the Subset Algorithm based on Response Variables~
The Hypothesis of equality using response variables
  Sepal.Length Sepal.Width Petal.Length Petal.Width is rejected
The Hypothesis of equality using response variables
  Sepal.Width Petal.Length Petal.Width is rejected
The Hypothesis of equality using response variables
  Sepal.Length Petal.Length Petal.Width is rejected
The Hypothesis of equality using response variables
  Sepal.Length Sepal.Width Petal.Width is rejected
The Hypothesis of equality using response variables
  Sepal.Length Sepal.Width Petal.Length is rejected
The Hypothesis of equality using response variables
  Petal.Length Petal.Width is rejected
The Hypothesis of equality using response variables
  Sepal.Width Petal.Width is rejected
The Hypothesis of equality using response variables
  Sepal.Width Petal.Length is rejected
The Hypothesis of equality using response variables
  Sepal.Length Petal.Width is rejected
The Hypothesis of equality using response variables
  Sepal.Length Petal.Length is rejected
The Hypothesis of equality using response variables
  Sepal.Length Sepal.Width is rejected
The Hypothesis of equality using response variables  Petal.Width is rejected
The Hypothesis of equality using response variables  Petal.Length is rejected
The Hypothesis of equality using response variables  Sepal.Width is rejected
The Hypothesis of equality using response variables  Sepal.Length is rejected
All appropriate subsets using response variables have been checked using a
  multiple testing procedure, which controls the maximum overall type I
  error rate at alpha= 0.05
```

The results are not surprising: Regarding the species (factor levels), all pairwise comparisons are significant. Regarding the variables, every one of the four variables alone shows a significant difference between the species, and so does every combination of variables. The output thus also illustrates the maximum number of tests being performed that are possible for the given design configuration (here: number of factor levels $a = 3$, number of variables $p = 4$). Namely, $2^a - a - 1 = 4$ tests are being performed for factor level combinations, and $2^p - 1 = 15$ tests for sets of response variables.

4. Conclusion

In the preceding sections, we have presented the R package **npmv** that enables researchers to make sense of multivariate data samples. The package performs valid nonparametric inference for the comparison of the samples and supplements it by appropriate graphical and numerical descriptive information. The package **npmv** has two main functions, `nonpartest` and `ssnonpartest`. The function `nonpartest` tests the global hypothesis and provides boxplots (for smaller data sets: dot-plots), as well as estimators of the nonparametric relative treatment effects. In case of overall significant results, the `ssnonpartest` function can be used to perform an all subset testing algorithm which maintains the familywise error rate and determines which variables or factor levels caused the significance. The results are comprehensively summarized in standard language that can be used directly in research papers applying the statistical methodology.

Future versions of the package will extend the methods presented to missing data and factorial designs. For both situations, the theory is still being developed.

References

Acion L, Peterson JJ, Temple S, Arndt S (2006). "Probabilistic Index: An Intuitive Non-Parametric Approach to Measuring the Size of Treatment Effects." *Statistics in Medicine*, **25**(4), 591–602. doi:10.1002/sim.2256.

Anderson E (1935). "The Irises of the Gaspe Peninsula." *Bulletin of the American Iris Society*, **59**, 2–5.

Bathke AC, Harrar SW (2008). "Nonparametric Methods in Multivariate Factorial Designs for Large Number of Factor Levels." *Journal of Statistical Planning and Inference*, **138**(3), 588–610. doi:10.1016/j.jspi.2006.11.004.

Bathke AC, Harrar SW, Ahmad MR (2009). "Some Contributions to the Analysis of Multivariate Data." *Biometrical Journal*, **51**(2), 285–303. doi:10.1002/bimj.200800196.

Bathke AC, Harrar SW, Madden LV (2008). "How to Compare Small Multivariate Samples Using Nonparametric Tests." *Computational Statistics & Data Analysis*, **52**(11), 4951–4965. doi:10.1016/j.csda.2008.04.006.

Brunner E, Domhof S, Langer F (2002). *Nonparametric Analysis of Longitudinal Data in Factorial Experiments*. John Wiley & Sons, New York.

Burchett W, Ellis A (2017). ***npmv**: Nonparametric Comparison of Multivariate Samples*. R package version 2.4.0, URL https://CRAN.R-project.org/package=npmv.

Fisher RA (1936). "The Use of Multiple Measurements in Taxonomic Problems." *The Annals of Eugenics*, **7**(2), 179–188. doi:10.1111/j.1469-1809.1936.tb02137.x.

Grabcanovic-Musija F, Obermayer A, Stoiber W, Krautgartner WD, Steinbacher P, Fuchs N, Bathke AC, Klappacher M, Studnicka M (2015). "Neutrophil Extracellular Trap (NET) Formation Characterises Stable and Exacerbated COPD and Correlates with Airflow Limitation." *Respiratory Research*, **16**(59). doi:10.1186/s12931-015-0221-7.

Harrar SW, Bathke AC (2008a). "Nonparametric Methods for Unbalanced Multivariate Data and Many Factor Levels." *Journal of Multivariate Analysis*, **99**(8), 1635–1664. doi:10.1016/j.jmva.2008.01.005.

Harrar SW, Bathke AC (2008b). "A Nonparametric Version of the Bartlett-Nanda-Pillai Multivariate Test. Asymptotics, Approximations, and Applications." *American Journal of Mathematical and Management Sciences*, **28**(3–4), 309–335. doi:10.1080/01966324.2008.10737731.

Hommel G (1985). "Multiple Vergleiche Mittels Rangtests – Alle Paarvergleiche." In *Neuere Verfahren der nichtparametrischen Statistik*, pp. 28–48. Springer-Verlag, Berlin.

Horst LE, Locke J, Krause CR, McMahon RW, Madden LV, Hoitink HAJ (2005). "Suppression of Botrytis Blight of Begonia by Trichoderma Hamatum 382 in Peat and Compost-Amended Potting Mixes." *Plant Disease*, **89**(11), 1195–1200. doi:10.1094/pd-89-1195.

Hothorn T, Hornik K, Van de Wiel MA, Zeileis A (2006). "A Lego System for Conditional Inference." *The American Statistician*, **60**(3), 257–263. doi:10.1198/000313006x118430.

Hothorn T, Hornik K, Van de Wiel MA, Zeileis A (2008). "Implementing a Class of Permutation Tests: The **coin** Package." *Journal of Statistical Software*, **28**(8), 1–23. doi:10.18637/jss.v028.i08.

Kieser M, Friede T, Gondan M (2013). "Assessment of Statistical Significance and Clinical Relevance." *Statistics in Medicine*, **32**(10), 1707–1719. doi:10.1002/sim.5634.

Konietschke F, Placzek M, Schaarschmidt F, Hothorn LA (2015). "**nparcomp**: An R Software Package for Nonparametric Multiple Comparisons and Simultaneous Confidence Intervals." *Journal of Statistical Software*, **64**(9), 1–17. doi:10.18637/jss.v064.i09.

Liu C, Bathke AC, Harrar SW (2011). "A Nonparametric Version of Wilks' Lambda – Asymptotic Results and Small Sample Approximations." *Statistics and Probability Letters*, **81**, 1502–1506. doi:10.1016/j.spl.2011.04.012.

Marcus R, Peritz E, Gabriel KR (1976). "On Closed Test Procedures with Special Reference to Ordered Analysis of Variance." *Biometrika*, **63**(3), 655–660. doi:10.1093/biomet/63.3.655.

Munzel U, Brunner E (2000a). "Nonparametric Methods in Multivariate Factorial Designs." *Journal of Statistical Planning and Inference*, **88**(1), 117–132. doi:10.1016/s0378-3758(99)00212-8.

Munzel U, Brunner E (2000b). "Nonparametric Tests in the Unbalanced Multivariate One-Way Design." *Biometrical Journal*, **42**(7), 837–854. doi:10.1002/1521-4036(200011)42:7<837::aid-bimj837>3.0.co;2-s.

Munzel U, Brunner E (2001). "Corrections of "Nonparametric Tests in the Unbalanced Multivariate One-Way Design"." *Biometrical Journal*, **43**(6), 791–792. doi:10.1002/1521-4036(200110)43:6<791::aid-bimj791>3.0.co;2-w.

Nardone R, Höller Y, Bathke AC, Höller P, Lochner P, Tezzon F, Trinka E, Brigo F (2014). "Subjective Memory Impairment and Cholinergic Transmission: A TMS Study." *Journal of Neural Transmission*, **122**(6), 873–876. doi:10.1007/s00702-014-1344-6.

Nardone R, Höller Y, Thomschewski A, Bathke AC, Ellis AR, Golaszewski SM, Brigo F, Trinka E (2015). "Assessment of Corticospinal Excitability after Traumatic Spinal Cord Injury Using MEP Recruitment Curves: A Preliminary TMS Study." *Spinal Cord*, **53**, 534–538. doi:10.1038/sc.2015.12.

Noguchi K, Gel YR, Brunner E, Konietschke F (2012). "**nparLD**: An R Software Package for the Nonparametric Analysis of Longitudinal Data in Factorial Experiments." *Journal of Statistical Software*, **50**(12), 1–23. doi:10.18637/jss.v050.i12.

Oksanen J, Blanchet FG, Kindt R, Legendre P, Minchin PR, O'Hara RB, Simpson GL, Solymos P, Stevens MHH, Wagner H (2016). ***vegan**: Community Ecology Package*. R package version 2.4-1, URL https://CRAN.R-project.org/package=vegan.

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rizzo ML, Szekely GJ (2016). ***energy**: E-Statistics: Multivariate Inference via the Energy of Data*. R package version 1.7-0, URL https://CRAN.R-project.org/package=energy.

Sonnemann E (2008). "General Solutions to Multiple Testing Problems." *Biometrical Journal*, **50**(5), 641–656. doi:10.1002/bimj.200810462. Translation of Sonnemann, E. (1982). Allgemeine Lösungen multipler Testprobleme. *EDV in Medizin und Biologie*, **13**(4), 120–128.

Székely GJ, Rizzo ML (2004). "Testing for Equal Distributions in High Dimension." *InterStat*, **2004**(11-5), 1–16. URL http://interstat.statjournals.net/YEAR/2004/abstracts/0411005.php.

Wheeler B (2016). ***lmPerm**: Permutation Tests for Linear Models*. R package version 2.1.0, URL https://CRAN.R-project.org/package=lmPerm.

Zeileis A, Croissant Y (2010). "Extended Model Formulas in R: Multiple Parts and Multiple Responses." *Journal of Statistical Software*, **34**(1), 1–13. doi:10.18637/jss.v034.i01.

A. Mathematical formulae

Define $R_{ij}^{(k)}$ as the rank (i.e., midrank) of $X_{ij}^{(k)}$ among all $N = \sum_{i=1}^{a} n_i$ random variables $X_{11}^{(k)}, \ldots, X_{a,n_a}^{(k)}$. Here, $X_{ij}^{(k)}$ is the observation on variable k at subject j in group i, and N is the total number of subjects (experimental units) in the study. We define the vectors $\mathbf{R}_{ij} = (R_{ij}^{(1)}, \ldots, R_{ij}^{(p)})^\top$ containing all the ranks of one multivariate observation, and the $p \times N$ matrix $\mathbf{R} = (\mathbf{R}_{11}, \ldots, \mathbf{R}_{1n_1}, \mathbf{R}_{21}, \ldots, \mathbf{R}_{an_a})$ containing the ranks for all variables and all observations. Here, a variable corresponds to a row, and a multivariate observation corresponds to a column of the matrix $\mathbf{R}$, as illustrated in Table 1 for the matrix of original observations.

Formally, define the sums of squares and cross-products

$$\mathbf{H}_1 = \frac{1}{a-1}\mathbf{R}\left(\bigoplus_{i=1}^{a} \frac{1}{n_i} J_{n_i} - \frac{1}{N} J_N\right)\mathbf{R}^\top , \quad \mathbf{G}_1 = \frac{1}{N-a}\mathbf{R}\left(\bigoplus_{i=1}^{a} P_{n_i}\right)\mathbf{R}^\top ,$$

$$\mathbf{H}_2 = \frac{1}{a-1}\mathbf{R}\left[\left(\bigoplus_{i=1}^{a} \frac{1}{n_i}\mathbf{1}_{n_i}\right) P_a \left(\bigoplus_{i=1}^{a} \frac{1}{n_i}\mathbf{1}_{n_i}^\top\right)\right]\mathbf{R}^\top , \quad \mathbf{G}_2 = \frac{1}{a}\mathbf{R}\left(\bigoplus_{i=1}^{a} \frac{1}{n_i(n_i-1)} P_{n_i}\right)\mathbf{R}^\top .$$

In this notation, the pair $(\mathbf{H}_1, \mathbf{G}_1)$ corresponds to a weighted means analysis, while the pair $(\mathbf{H}_2, \mathbf{G}_2)$ uses unweighted means. In a balanced design with $n_i \equiv n$, $i = 1, \ldots, a$, $\mathbf{H}_1 = n \cdot \mathbf{H}_2$ and $\mathbf{G}_1 = n \cdot \mathbf{G}_2$. Therefore, in a balanced design, both pairs lead to the same test statistic. Extensive simulation studies (see, e.g., Bathke *et al.* 2008) have not shown any systematic advantages of one pair over the other. Considering also that well-planned studies typically strive for experimental designs that are close to balanced, the difference between using the matrix pair $(\mathbf{H}_1, \mathbf{G}_1)$ or $(\mathbf{H}_2, \mathbf{G}_2)$ appears to be negligible in most practical applications. Historically, many multivariate test statistics have been defined using $(\mathbf{H}_1, \mathbf{G}_1)$, while the ANOVA-type statistic was first introduced using $(\mathbf{H}_2, \mathbf{G}_2)$ (see, e.g., Munzel and Brunner 2000a,b, 2001).

We consider four types of test statistics: ANOVA type, Wilks' Lambda type, Lawley Hotelling type, and Bartlett Nanda Pillai type. For each of the four, a moment-based finite sample approximation based on quantiles of the F-distribution is derived. Additionally, the package calculates permutation (randomization) p values.

ANOVA type statistic

The ANOVA type statistic is defined as $T_A = \mathrm{tr}(\mathbf{H}_2)/\mathrm{tr}(\mathbf{G}_2)$. The distribution of T_A is approximated by an F-distribution with estimated degrees of freedom $\hat{f}_1$ and $\hat{f}_2$, where

$$\hat{f}_1 = \frac{\mathrm{tr}(\mathbf{G}_2)^2}{\mathrm{tr}(\mathbf{G}_2{}^2)} \qquad \text{and} \qquad \hat{f}_2 = \frac{a^2}{(a-1)\sum_{i=1}^{a}\frac{1}{n_i-1}} \cdot \hat{f}_1.$$

Wilks' Lambda type

Wilks' Lambda type statistic is defined as

$$\lambda = \frac{\det[(N-a)\cdot\mathbf{G}_1]}{\det[(N-a)\cdot\mathbf{G}_1 + (a-1)\cdot\mathbf{H}_1]} .$$

The sampling distribution of

$$F_\lambda = [(1-\lambda^{1/t})/(\lambda^{1/t})](df_2/df_1)$$

is approximated by an F-distribution where $df_1 = p(a-1)$, $df_2 = rt - (p(a-1)-2)/2$, and $r = (N-a) - (p-(a-1)+1)/2$. If $p(a-1) = 2$, then $t = 1$, else $t = \sqrt{\frac{p^2(a-1)^2-4}{p^2+(a-1)^2-5}}$.

Lawley Hotelling type

The Lawley Hotelling type statistic is calculated as

$$U = \mathrm{tr}[(a-1)\mathbf{H}_1((N-a)\mathbf{G}_1)^{-1}] \ .$$

The distribution of U is approximated by $g \times F_{K,D}$, a "stretched" F-distribution with degrees of freedom K and D, where $K = p(a-1)$, $D = 4 + \frac{K+2}{B-1}$, $B = \frac{(N-p-2)(N-a-1)}{(N-a-p)(N-a-p-3)}$, and $g = \frac{p(a-1)(D-2)}{(N-a-p-1)D}$.

Bartlett Nanda Pillai type

The Bartlett Nanda Pillai type statistic is defined as

$$V = \mathrm{tr}\{(a-1)\mathbf{H}_1[(a-1)H_1 + (N-a)\mathbf{G}_1]^{-1}\} \ .$$

The distribution of $\frac{(V/\gamma)/\nu_1}{(1-V/\gamma)/\nu_2}$ is approximated using an F-distribution with degrees of freedom ν_1 and ν_2, where

$$\begin{aligned}
\gamma &= \min(a-1,p) \\
\nu_1 &= \frac{p(a-1)}{\gamma(N-1)}\Big[\frac{\gamma(N-a+\gamma-p)(N+2)(N-1)}{(N-a)(N-p)} - 2\Big] \\
\nu_2 &= \frac{N-a+\gamma-p}{N}\Big[\frac{\gamma(N-a+\gamma-p)(N-1)}{(N-a)(N-p)} - 2\Big].
\end{aligned}$$

See Bathke and Harrar (2008); Harrar and Bathke (2008a,b); Bathke *et al.* (2008, 2009); Liu *et al.* (2011) for the derivations of asymptotic results and small sample approximations for each of the test statistics presented above.

Affiliation:

Arne C. Bathke
Fachbereich Mathematik
Universität Salzburg
A-5020 Salzburg, Austria
E-mail: Arne.Bathke@sbg.ac.at

Journal of Statistical Software http://www.jstatsoft.org/
published by the Foundation for Open Access Statistics http://www.foastat.org/
January 2017, Volume 76, Issue 4 *Submitted:* 2013-07-30
doi:10.18637/jss.v076.i04 *Accepted:* 2015-11-02