The welchADF Package for Robust Hypothesis Testing in Unbalanced Multivariate Mixed Models with Heteroscedastic and Non-normal Data

by Pablo J. Villacorta

Abstract A new R package is presented for dealing with non-normality and variance heterogeneity of sample data when conducting hypothesis tests of main effects and interactions in mixed models. The proposal departs from an existing SAS program which implements Johansen's general formulation of Welch-James's statistic with approximate degrees of freedom, which makes it suitable for testing any linear hypothesis concerning cell means in univariate and multivariate mixed model designs when the data pose non-normality and non-homogeneous variance. Improved type I error rate control is obtained using bootstrapping for calculating an empirical critical value, whereas robustness against non-normality is achieved through trimmed means and Winsorized variances. A wrapper function eases the application of the test in common situations, such as performing omnibus tests on all effects and interactions, pairwise contrasts, and tetrad contrasts of two-way interactions. The package is demonstrated in several problems including unbalanced univariate and multivariate designs.

Introduction

The problem of testing for mean equality between several groups can be accomplished using classical techniques such as Student's t test, when only two groups are compared, or ANOVA when more than two groups are involved. Both of them have been widely applied in the past in a number of areas ranging from ecology and biology to psychology, social sciences and medicine (Levin, 1997; Coates and McKenzie-Mohr, 2010), although their use tends to decrease for the reasons mentioned next.

In order for these approaches to work well, data must satisfy three conditions, namely independence, normality and homoscedasticity of the errors. While ANOVA is known to be robust to small departures from normality, homogeneity of population variances is crucial as concluded by simulation studies in which these methods have been found to exhibit type I error rates oscillating from too conservative to extremely liberal, specially in unbalanced designs leading to very heterogeneous cell variances. The behaviour depends on the relation between the variance of the cell with the fewest observations and the number of observations contained in it (Milligan et al., 1987). Indeed, data from a number of experiments conducted in the aforementioned research fields often exhibit non-homogeneous variances. This is not a problem for one-way designs since the built-in function `oneway.test` in package **stats** is able to account for different variances. Unfortunately, real-world studies usually require more complex designs, like the typical mixed between x within-subjects designs for clinical trials involving either animals or persons. For both reasons, the application of simple ANOVA in serious analyses of experimental data is nowadays not very common.

A distinction should be made here on what homogeneity of variances means depending on the design being considered (Lix and Keselman, 1995). In univariate settings with between-subjects factors only, all the cell variances should be equal, while in multivariate settings, it refers to the equality of the population covariance matrices across all cells. In mixed designs with at least one between-subjects factor, a set of orthonormalized contrasts on the repeated measures must have common covariance matrices (sphericity assumption), and all those matrices must be equal across all cells of the between-subject factors. When both conditions are met, it is said that multisample sphericity holds (Huynh, 1978).

A number of alternatives have been proposed to overcome the parametric restrictions (Higgins, 2003). Traditionally the most widely used choices have been nonparametric tests such as Mann-Whitney-Wilcoxon rank sum test and Kruskal test for two groups, or Wilcoxon signed-rank test and Friedman test (for which paired versions exist) for more than two groups. However, they cannot handle multiple factors or interactions. Generalized Linear Models can deal with multiple factors and interactions with non-normal data, but require specifying the link function and are unable to handle repeated measures. Since our study focuses on the most general models, i.e. those with between-subjects and within-subjects interactions, the aforementioned tests are not useful. The generalization of Mixed Models, namely Generalized Linear Mixed Models (Bolker et al., 2009), do constitute a valid alternative. They present some drawbacks, though, such as being complex to apply and interpret, not very widely available, and requiring a particular treatment for each problem since a suitable link

function must be supplied in each case, which is not always possible.

Two surveys on nonparametric techniques in experimental design can be found in Sawilowsky (1990); Salazar-Alvarez et al. (2014). In these works, several rank transformation variants are emphasized, as they constitute the most widely used nonparametric approach for detecting interactions Conover (2012). Among them, the Aligned Rank Transform (Higgins and Tashtoush, 1994), for which an implementation in R has been made available in package **ART** (Villacorta, 2015), is one of the best performing for this task, keeping type I error rates close to the theoretical significance level while preserving good power. Although it has been applied to a split-plot design (one between- and one within-subjects factor) in Beasley (2002), showing good type I error rates and power, it lacks a general unified formulation for mixed models with any number of between- and within-subjects factors that also works in unbalanced and multivariate settings. Erce-Hurn and Mirosevich (2008); Ruscio and Roche (2012) constitute two more broad surveys (the latter dealing with variance heterogeneity in depth) covering both classical nonparametric methods and recent research efforts like the one implemented here, which is cited in both works.

Some other valid alternatives for nonparametric analysis of any mixed model are the Improved General Approximation (IGA), the generalization of Welch-James (WJ) test statistic, the Kenward-Roger correction with mixed models (KR), and the modified Brown and Forsythe (MBF) procedure. They all do a correction for the degrees of freedom to account for heterogeneous variances, hence the name ADF for *approximate degrees of freedom*. The IGA (Huynh, 1978) was specifically developed to account for multisample sphericity violations in repeated measures designs by adjusting the critical value, and was generalized to any mixed model by Algina (1997). Similarly, Welch's non-pooled statistic with approximate degrees of freedom (ADF) (Welch, 1951) was also conceived for this purpose, using the sample data to estimate the error degrees of freedom. Several non-pooled ADF statistics have been proposed later but all can be derived from the general matrix formulation of Welch's statistic given by Lix and Keselman (1995) based on Johansen (1980), which makes it applicable for univariate and multivariate mixed models with an arbitrary number of effects. Lix and Keselman (1995) shows how the same statistic can be employed in different models for both omnibus contrasts (testing whether a given effect is significant or not) and pairwise comparisons for a given effect (for every pair of categories of a given effect, testing whether the response is significantly different for one of the categories against one another). Generally, the IGA and WJ statistic perform similarly; however a slight advantage favorable to WJ has been reported by Algina (1997) in some contexts. The WJ ADF approach has been successfully tested in nonparametric analysis of a variety of mixed models; see Keselman et al. (2003) and references therein. The KR correction for the degrees of freedom can be implemented on top of a conventional mixed model and performs similarly to the MBF procedure with slight advantage for the latter (Vallejo and Livacic-Rojas, 2005). Both yield reasonably good results. With respect to the comparison between the MBF and the generalized WJ statistic, there is no consensus about the results. In most conditions they perform similarly, but some authors state that MBF is better at detecting interaction effects when the number of subjects is not high enough (Vallejo et al., 2001, 2006). However, to the best of our knowledge, there is no general formulation of MBF for any mixed model with an arbitrary number of between- x within-subjects factors, although it has been tested in split-plot designs (Vallejo et al., 2006) (which are probably the most common design in medicine and particularly in psychological studies) and factorial designs (Vallejo et al., 2008).

Software available for robust testing of mixed models not meeting parametric assumptions

Wilcox (2012) constitutes an important source dealing with robust estimation. The book is accompanied by an R package called **WRS** [1] that implements all the methods reviewed in the book, including the Welch-James test following Johansen's approach with robust mean estimators described in sections 7.2, 8.6 and 8.7 which our package **welchADF** also implements. The functions described in those sections, `bwtrim`,`t1way`,`t1waybt`,`t2way`,`t2way`,`t3way`, include the most common designs such as one, two and three-way and split-plot (one between x one within subjects designs) also with bootstrapping and trimming. Although not included in **WRS2**, the original **WRS** exposes functions `bbw-`,`bww-` for between x between x within, and between x within x within subjects designs, and their trimmed and bootstrap versions. While useful, they do not provide a uniform, easy-to-use interface in a single function valid for any mixed design.

There exists a CRAN task force on robust statistical methods[2]. Unfortunately, none of the packages mentioned there implements the aforementioned approaches. Nevertheless, packages **robustbase** (Maechler et al., 2016), **robust** (Wang et al., 2017) (by the authors of Maronna et al. (2006)) and function

[1] Not on CRAN but in `http://github.com/nicebread/WRS`. A less-comprehensive but more user-friendly version can be found in package **WRS2** (Mair and Wilcox, 2017).

[2] `http://cran.r-project.org/web/views/Robust.html`

`rlmer` in package **robustlmm** (Koller, 2017) are some remarkable efforts. The latter implements the techniques proposed in Koller (2013) for linear mixed models on a basis of a parametric model having some contaminated data (Koller, 2016). However, it does not focus on testing inherently heterogeneous and non-normal data. Package **nlme** (nlm, 2017) provides a way of capturing and modeling variance heterogeneity in mixed models through the argument `weights` of function `lme`, which can be set to different covariance matrix structures that are fitted from the data. The well-known package **lme4** (Bates et al., 2016, 2015) is also a good choice for dealing with non-normal models in presence of within-subjects effects (called generalized linear mixed models as an extension to generalized linear models where the user can specify the probabilistic model to be used). Package **glmmADMB** [3] (Skaug et al., 2016; Fournier et al., 2012) has a similar purpose.

SAS (SAS Institute, 2011) implementations of the WJ statistic, MBF statistic (split-plot and factorial designs only) and IGA do exist; see Keselman et al. (2003); Vallejo et al. (2006); Algina (1997) respectively. While SAS is still widely used mainly in social and biomedical sciences, it is proprietary software. The same applies to the ERP-PCA software (Dien, 2010) written in Matlab. Interestingly, ERP-PCA incorporates a Matlab translation of the SAS code described in Keselman et al. (2003) for the WJ statistic. For these reasons, together with the fast expansion of R and open-source statistical software in general -closely related to the growing interest in reproducible research- among researchers of many different disciplines, we consider the R package introduced here a useful effort.

Main contributions

The present work describes an R package called **welchADF** (Villacorta, 2017) that implements Johansen's formulation of the Welch-James test, with two additional improvements: first, the use of trimmed means and Winsorized variances to deal with non-normality, and second, the use of bootstrap for calculating an empirical critical value for achieving better type I error control. Both aspects are mentioned in Wilcox et al. (1998) and implemented in Keselman et al. (2003). Trimmed means and Winsorized variances are being used in medical and behavioral research in the last years; see Müller et al. (2011); Aronoff et al. (2011); Ryzin et al. (2011).

The core of our code is an R translation of the SAS program[4] described in Keselman et al. (2003). A new wrapper function has been built on top of it that poses the following benefits:

- It can be applied to univariate and multivariate mixed models with an arbitrary number of within- and between-subjects effects.
- It simplifies some common tasks such as performing omnibus tests on effects or interactions, multiple pairwise comparisons on the levels of one factor, and tetrad contrasts. All of these can be done without indicating the contrast matrices, which are automatically formed by the program depending on the kind of test required and the number of levels found in each factor.
- It provides a more natural and uniform data input mechanism through data frames that do not depend on the model specified. In the original SAS code, the input data had to be carefully arranged in matrices whose shape had to mirror the experimental design being analyzed in each problem, which can be error-prone.
- It integrates with other similar packages of the R ecosystem through a formula interface and also provides additional interfaces that accept model objects returned by some commonly used functions such as `stats::lm`, `lme4::lmer` and `stats::aov`.
- It enables selecting one among several built-in p-value correction methods when performing multiple pairwise comparisons.

There are several reasons that justify an R implementation of this particular test:

- The generalized WJ test described in Lix and Keselman (1995); Keselman et al. (2003) has good theoretical properties and has proven successful in controlling type I error rate while preserving high power. Moreover, the use of trimmed means and Winsorized variances can protect both against skewness and outliers in the data, as noted by Keselman et al. (2008). The percentage of trimming can be adjusted to deal with higher ratios of outliers and skewness. The statistic is also able to cope with heterogeneous variances, which commonly (but not only) arise when having very different cell sample sizes. In this sense, one should notice the sample size requirement stated right after this list.
- These two works have received a number of citations, and the approach explained there is being used in current research in different fields such as medicine (Dien et al., 2008; Dien, 2010), psychology (Müller et al., 2011; Kayser et al., 2014; Huang and Jun, 2015) and behavioral research (Symes et al., 2010), just to cite a few.

[3] http://glmmadmb.r-forge.r-project.org/

[4] Available at http://homepage.usask.ca/~lml321/Program.pdf

- The function interface is simple and the test can be used in a straightforward way for the most common tasks. This may contribute positively towards its adoption by the research community, specially by researchers with little expertise in statistics, but with an understanding of the importance of applying suitable, robust techniques when parametric conditions are not met.
- Despite the existence of different alternatives explained before, some of them also implemented in R, these are either not applicable to multivariate mixed models with heterogeneous variance or non-normal data, or are generally complex and more difficult to use.

The WJ approach also has some disadvantages. The first one is the sample size needed to assure an effective control of type I error under some (somewhat extreme) circumstances, specially in repeated-measures designs, like when the cell with the fewest subjects presents the largest variance (i.e. cell size and cell variance are negatively paired). In general, the number of subjects of the smallest cell should be four or five times greater than the number of repeated measures minus one, and sometimes even more when testing an interaction. However, when combined with trimmed means, robustness is increased and some of these problems are mitigated as a much smaller number of subjects are required (Keselman et al., 2000). The second drawback is that **welchADF** is only applicable to categorical predictors. In case the design has numeric predictors, the reader may try the packages mentioned in the preceding section, as well as **gamm4** (Wood and Scheipl, 2017) and **mgcv** (Wood, 2017) for fitting generalized additive mixed models.

Finally, the application of bootstrap to contaminated data has been extensively studied and even questioned by some authors in the past (Singh, 1998), specially regarding numerical instability: some bootstrap samples that intervene in the computation of the final bootstrapped estimate may contain a higher proportion of outliers than the general dataset, and therefore be too heavily influenced by them (Salibián-Barrera et al., 2008). Singh (1998) proposed using bootstrapping with Winsorized observations, which is what we do in this package when enabling trimming and bootstrapping at the same time, as it provides some additional benefits. Salibián-Barrera et al. (2008) introduce a fast and robust bootstrap (FRB) method that improves classical bootstrapping. Although it has not been incorporated to **welchADF**, it may be done in the future.

The remainder of this contribution is structured as follows. In first place, the mathematical background is briefly reviewed. In second place, we present the function exposed by the package and explain its arguments, together with some issues regarding the arrangement of the data. In third place, we address three case studies, namely a univariate one-way between-subjects design, a two-way factorial design, a mixed design, and a doubly multivariate design analyzed as a multivariate mixed design. Finally, we present some conclusions and further work.

The Welch-James ADF test statistic

Here we summarize the theoretical background given in Lix and Keselman (1995); Keselman et al. (2003). Following the General Linear Model,

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\xi} \tag{1}$$

where $\mathbf{Y}$ is an $N \times p$ matrix of observations on p dependent variables or p repeated measures, N is the sample size, $\mathbf{X}$ is the design matrix (formed by zeros and ones, such that $\text{rank}(\mathbf{X}) = r$ with r being the number of different groups or cells[5]), $\boldsymbol{\beta}$ is an $r \times p$ matrix of non random (unknown) parameters (population means) and $\boldsymbol{\xi}$ is an $N \times p$ matrix of random error components.

If we denote by $\mathbf{Y}_j (j = 1, ..., r)$ the submatrix of $\mathbf{Y}$ that contains the observations of the n_j subjects of the j-th cell, the original parametric model assumes $\mathbf{Y}_j \sim \mathcal{N}(\boldsymbol{\beta}_j, \boldsymbol{\Sigma}_j)$ where $\boldsymbol{\beta}_j = (\mu_{j1}, ..., \mu_{jp})$ is the j-th row of matrix $\boldsymbol{\beta}$, and $\boldsymbol{\Sigma}_j \neq \boldsymbol{\Sigma}_{j'}$ when $j \neq j'$.

We proceed to explain how population means and variances are estimated. The matrix of population means can be estimated by the usual least-squares approach (Eq. 2) or by using robust estimation techniques such as trimming, discussed later. Now, let $\mathbf{X}_j (j = 1, ..., r)$ be the j-th column of $\mathbf{X}$, composed of zeros and ones, and let $\mathbf{1}_p$ be a $p \times 1$ vector of ones. Define $\mathbf{Y}_j = \mathbf{Y} \circ (\mathbf{X}_j \mathbf{1}_p^T)$ as the $N \times p$ matrix that results of the Hadamard (i.e. element-wise) product between matrices $\mathbf{Y}$ and $\mathbf{X}_j \mathbf{1}_p^T$. Then, the following expressions are used to estimate population means and variances:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \tag{2}$$

$$\hat{\boldsymbol{\Sigma}}_j = \frac{(\mathbf{Y}_j - \mathbf{X}_j\hat{\boldsymbol{\beta}}_j)^T(\mathbf{Y}_j - \mathbf{X}_j\hat{\boldsymbol{\beta}}_j)}{n_j - 1} \tag{3}$$

[5]This does not entail that one-way designs are the only possible. The formulation is valid also for several factor designs.

The matrix formulation of the WJ statistic always tests the following general linear hypothesis:

$$H_0 : \mathbf{R}\boldsymbol{\mu} = \mathbf{0} \tag{4}$$

with $\mathbf{R} = \mathbf{C} \otimes \mathbf{U}^T$. In this expression, $\otimes$ is the Kronecker product, $\mathbf{C}$ is a $df_C \times r$ matrix that indicates the contrasts on the between-subjects effects, so that rank($\mathbf{C}$) $= df_C \leq r$, and $\mathbf{U}$ is a $p \times df_U$ matrix that indicates the contrasts on the within-subjects effects, so that rank($\mathbf{U}$) = $df_U \leq p$. Therefore $\mathbf{R}$ has $df_C df_U$ rows and rp columns. Note that $\boldsymbol{\mu} = \text{vec}(\boldsymbol{\beta}^T) = (\boldsymbol{\beta}_1, ...\boldsymbol{\beta}_r)^T$, which is a column vector with rp elements obtained by stacking the columns of $\boldsymbol{\beta}^T$, one on top of another. Matrices $\mathbf{C}$ and $\mathbf{U}$ determine the type of contrast being performed, be it an omnibus contrast, a pairwise contrast, etc. We provide details about the structure of both matrices in the general case in Section 2.2.1.

The generalized Welch-James test statistic presented by Johansen in Johansen (1980) is

$$\begin{aligned} T_{\text{WJ}} &= (\mathbf{R}\hat{\boldsymbol{\mu}})^T (\mathbf{R}\hat{\boldsymbol{\Sigma}}\mathbf{R}^T)^{-1} (\mathbf{R}\hat{\boldsymbol{\mu}}) \\ \hat{\boldsymbol{\Sigma}} &= \text{diag}(\hat{\boldsymbol{\Sigma}}_1/n_1, ..., \hat{\boldsymbol{\Sigma}}_r/n_r) \end{aligned} \tag{5}$$

where $\hat{\boldsymbol{\mu}}$ is an estimate of $\boldsymbol{\mu}$ (either with LS or any other technique), and $\hat{\boldsymbol{\Sigma}}$ is a block diagonal matrix whose blocks are $\hat{\boldsymbol{\Sigma}}_j/n_j$. It is known that T_{WJ}/c approximately follows an $F(\nu_1; \nu_2)$ where

$$\begin{aligned} \nu_1 &= df_C df_U \quad ; \quad \nu_2 = \nu_1(\nu_1+2)/(3A) \quad ; \quad c = \nu_1 + 2A - (6A)/(\nu_1+2) \\ A &= \frac{1}{2}\sum_{j=1}^{r} \frac{\text{tr}\left[\hat{\boldsymbol{\Sigma}}\mathbf{R}^T(\mathbf{R}\hat{\boldsymbol{\Sigma}}\mathbf{R}^T)^{-1}\mathbf{R}\mathbf{Q}_j\right]^2 + \left(\text{tr}\left[\hat{\boldsymbol{\Sigma}}\mathbf{R}^T(\mathbf{R}\hat{\boldsymbol{\Sigma}}\mathbf{R}^T)^{-1}\mathbf{R}\mathbf{Q}_j\right]\right)^2}{n_j - 1} \end{aligned}$$

where tr is the trace of an square matrix (sum of the elements on the main diagonal), and $\mathbf{Q_j}$ is an $rp \times rp$ matrix associated with X_j in which the (s,t)-th diagonal block of $\mathbf{Q}_j = \mathbf{I}_p$ when $s = t = j$ and is $\mathbf{0}$ otherwise.

In a between-subjects design (no within-subjects factors), $\mathbf{U}$ must be set to I_p where p is the number of dependent variables (in univariate designs, it reduces to $\mathbf{U} = 1$). In a within-subjects design (no between-subjects factors), $\mathbf{C}$ must be set to 1 both in the univariate and multivariate cases.

Structure of the contrast matrices

The preceding formulation of the model is valid for any type of contrasts. Most often the user may want to perform two types of contrasts, namely omnibus contrasts to check whether a given effect or interaction is statistically significant, and in case it is, post-hoc pairwise contrasts on one effect or interaction to check whether the response associated to some of the levels of that factor or interaction is statistically different than the responses associated to other levels of the factor.

Omnibus contrasts This test is aimed at checking whether the level adopted by a given variable of interest (effect or interaction) has an influence over the response variable. In the simplest case, consider a one-way design, either univariate or multivariate, whose single factor A has a different levels. Then $\mathbf{C}$ would be an $(a-1) \times a$ matrix specifying linearly independent contrasts between the levels. We will call this matrix $\mathbf{C}_A$ (capital A) because it is the matrix we use to conduct an omnibus test on effect A. If for instance, $a = 3$, we would have

$$\mathbf{C}_A = \begin{bmatrix} 1 & -1 & 0 \\ 1 & 0 & -1 \end{bmatrix} \quad ; \quad \mathbf{U} = \mathbf{I}_p$$

Algina and Olejnik (1984) provide a general formulation to compose matrix $\mathbf{C}$ for omnibus tests on factorial designs with several between-subjects factors. The same idea, slightly modified, is also valid to compose matrix $\mathbf{U}$ in designs with several within-subjects factors. These procedures have been implemented in our **welchADF** package. Let C_a (non-capital a) be the $(a-1) \times a$ matrix of linear contrasts associated to the a levels of effect A. Its rows define $a-1$ linearly independent contrasts between the levels of A. When the design has no more factors than A, then $\mathbf{C}_A = C_a$ as in the case above. However, when more than one between-subjects factor exists, then $\mathbf{C}$ has to be properly set according to the factor to which we want to apply an omnibus test. In those cases, $\mathbf{C}$ is the Kronecker product of one matrix (or vector) per factor existing in the design, as follows.

Assume a between-subjects design with four effects A, B, C, D. Let $\mathbf{1}_a^T$ denote a $(1 \times a)$ vector of ones. If we want to perform an omnibus contrast on a main effect, say B, then $\mathbf{C} = \mathbf{C}_B = \mathbf{1}_a^T \otimes C_b \otimes \mathbf{1}_c^T \otimes \mathbf{1}_d^T$. In other words, if the factor matches the effect being tested, the corresponding contrast matrix appears in the product; otherwise, a vector of ones appears. The same applies to interactions. If we want to test the $B \times D$ interaction, for instance, we would have $\mathbf{C} = \mathbf{C}_{BD} = \mathbf{1}_a^T \otimes C_b \otimes \mathbf{1}_c^T \otimes C_d$. Since

factors B and D are involved in the interaction BD being tested, their contrast matrices appear in the product, while a vector of ones appears in the positions of the remaining factors of the design.

For within-subjects designs a similar rule applies. Assume that now A is a within-subjects factor, and let $U_a = C_a^T$, so that the columns of U_a define $a-1$ linearly independent contrasts between the levels of A. In a within-subjects design with several factors, the transposes of the U contrast matrices of those factors involved in the effect or interaction being tested appear in the Kronecker product; otherwise, a row vector of ones is used in that place as explained before. If the design is multivariate with p dependent response variables, an additional factor $\mathbf{I}_p$ always appears in the last place of the product. At the end, the result of the Kronecker product must be transposed again to obtain $\mathbf{U}$. For example, in a two-factor within-subjects multivariate design, the $\mathbf{U}$ matrices for conducting ominbus tests on effects A, B and the interaction AB would be, respectively, $\mathbf{U}_A = (U_a^T \otimes \mathbf{1}_b^T \otimes \mathbf{I}_p)^T$, $\mathbf{U}_B = (\mathbf{1}_a^T \otimes U_b^T \otimes \mathbf{I}_p)^T$, and $\mathbf{U}_{AB} = (U_a^T \otimes U_b^T \otimes \mathbf{I}_p)^T$.

In case we want to test a main effect (either a between- or a within-subjects effect) in a design containing both between- and within-subjects factors, the same rules apply: $\mathbf{C}$ and $\mathbf{U}$ are composed separately, and one of them will (for sure) be the result of the Kronecker product of vectors of ones only (including $\mathbf{I}_p$ as well when constructing $\mathbf{U}$ if the design is multivariate). Finally, if we want to test an interaction involving one or more between-subjects factors and one or more within-subjects, $\mathbf{C}$ must be formed as if we were testing only the between-subjects factors involved in the mixed interaction, and $\mathbf{U}$ as if testing the within-subjects factors, following the rules explained above.

Pairwise contrasts Now for a given effect, we are aimed at testing for every pair of categories of the effect whether the response is significantly different for one of the categories against one another. The procedure is similar to the omnibus contrasts. The only difference is that contrast matrices C_a and U_a associated to an effect A are replaced by contrast vectors, as follows. When testing for significant differences between factor levels j and j' of an effect A, either C_a is replaced by a row vector $\mathbf{c}_{jj'}$ if A is a between-subjects factor, or U_a is replaced by a column vectors $\mathbf{u}_{jj'}$ if A is a within-subjects factor. In a pairwise contrast vector, all positions are set to 0 except for those corresponding to the factor levels j and j' being tested, which are set to 1 and -1 respectively. These vectors are then used in the corresponding positions of the Kronecker products described before. For probing interactions (once the omnibus test on such interaction proved significant), tetrad contrasts have been implemented in our package. The null hypothesis being tested in a tetrad contrast involving two factors A and B, from which the interaction between levels j and j' from A, and k and k' from B is being tested, can be written as

$$H_{jj';kk'} : (\mu_{jk} - \mu_{jk'}) - (\mu_{j'k} - \mu_{j'k'}) = 0 \tag{6}$$

Trimmed means and Winsorized variances

Trimmed means help mitigate the effects of non-normality. When least-squares means are substituted by trimmed means, the null hypotheses being tested are the equality of population trimmed means: $\mathbf{R}\boldsymbol{\mu}^{(t)} = \mathbf{0}$. Let $Y_{(1)j} \leq Y_{(2)j} \leq ... \leq Y_{(n_j)j}$ be the sorted observations of the j-th group, and $g_j = \lceil \gamma n_j \rceil$ with γ being the proportion of observations to be trimmed in each tail of the distribution. Therefore the sample size for the j-th group becomes $h_j = n_j - 2g_j$, and its sample trimmed mean is computed by averaging the h_j central observations of that group:

$$\hat{\mu}_j^{(t)} = \frac{1}{h_j} \sum_{k=g_j+1}^{n_j-g_j} Y_{(k)j} \tag{7}$$

Some authors suggest using 20 % trimming.

The sample Winsorized mean is a similar measure that is computed by replacing all observations smaller than $Y_{(g_j+1)j}$ (i.e. the 20th percentile) by that value, and those larger than $Y_{(n_j-g_j)j}$ (i.e. the 80th percentile) by that value, and then averaging over all the (modified) observations. In the j-th group:

$$\hat{\mu}_j^{(W)} = \frac{1}{n_j} \sum_{i=1}^{n_j} X_{ij}, \quad \text{where} \quad X_{ij} = \begin{cases} Y_{(g_j+1)j} & \text{if} \quad Y_{ij} \leq Y_{(g_j+1)j} \\ Y_{ij} & \text{if} \quad Y_{(g_j+1)j} < Y_{ij} < Y_{(n_j-g_j)j} \\ Y_{(n_j-g_j)j} & \text{if} \quad Y_{ij} \geq Y_{(n_j-g_j)j} \end{cases} \tag{8}$$

This measure is required to compute the sample Winsorized variance:

$$\hat{\sigma}_j^{2(W)} = \frac{1}{n_j - 1} \sum_{i=1}^{n_j} (X_{ij} - \hat{\mu}_j^{(W)})^2 \tag{9}$$

Therefore, in order to compute the trimmed version of the WJ statistic, $T_{WJ}^{(t)}$, trimmed means replace least-squares means, Winsorized variances replace least-squares variances, and the new sample sizes h_j replace the original n_j in all groups. Past studies have found the trimmed version of the WJ statistic to be more robust and provide better type I error control to non-normality, also in more complex designs; see Keselman et al. (2000) and references therein.

Bootstrapping to obtain an empirical critical value

After computing the cell trimmed means, let $C_{ij} = Y_{ij} - \hat{\mu}_j^{(t)}$, i.e. the C_{ij} are shifted observations so that the null hypothesis of equal trimmed means is true in the sample. Repeat B times the next two steps (B is the user-supplied number of bootstrap simulations): (i) for each cell, generate a bootstrap sample of size n_j by sampling with replacement from the original sample. When the bootstrap samples of all the cells are put together, an N-sample bootstrap dataset is obtained; (ii) compute the value $F^{*(t)} = T_{WJ}^{(t)}/c$ on the bootstrap dataset. Sort the B values obtained in ascending order, $F_{(1)}^{*(t)} \leq ... \leq F_{(B)}^{*(t)}$. An estimate of an appropriate critical value is $F_{(a)}^{*(t)}$ where $a = (1-\alpha)B$ rounded to the nearest integer. This critical value should be compared with $T_{WJ}^{(t)}/c$ computed on the original data. The null hypothesis $\mathbf{R}\boldsymbol{\mu}^{(t)} = \mathbf{0}$ of trimmed means equality will be rejected if $T_{WJ}^{(t)}/c \geq F_{(a)}^{*(t)}$.

The process explained before applies to omnibus contrasts. For focused contrasts such as pairwise tests on marginal means, the same idea with minor modifications is used. For more details as well as a generalization to other designs, see Keselman et al. (2003) and references therein.

Effect size and confidence intervals

As stated in Keselman et al. (2008); APA (2013), it is now widely recommended to report an estimate of the effect size when performing a hypothesis test. Many different measures exist for this purpose, although few of them are valid for non-homogeneous variances. The approach implemented in our package was proposed in Keselman et al. (2008) and has the following formulation for the case of two groups:

$$\hat{\delta}_j^{(R)} = \eta \frac{\hat{\mu}_2^{(t)} - \hat{\mu}_1^{(t)}}{\hat{\sigma}_j^{(W)}} \tag{10}$$

Note this approach uses trimmed means and Winsorized standard deviation and for that reason, it is robust to non-normality. Factor η stands for a scaling factor of the Winsorized standard deviation. In case 20 % trimming is used (as recommended), $\eta = .642$ which is the Winsorized standard deviation for a 20% trimmed standard normal distribution. In our code, η is computed according to the percentage of trimming indicated by the user. For building a robust CI around this value, a percentile bootstrap method is run to determine the empirical bounds of the interval as recommended in Keselman et al. (2008).

An R implementation of the WJ statistic

The WJ test is implemented in our package as an S3 generic called `welchADF.test`. The name has been chosen to be compliant with other existing tests such as `t.test`,`wilcox.test`, etc. The function receives parameters to modulate its behaviour, such as the type of contrast to be performed (omnibus or pairwise), whether trimming should be employed or not, and if employed, the percentage of data to be trimmed at each side, and whether bootstrapping should be used or not. The default S3 method expects a `data.frame` in the `formula` parameter, but additional S3 methods are provided for classes `formula`, `lm`, `lmer` and `aov`, which allow our package to integrate well with other linear models functions, as described later in this section.

The prototype of the S3 default method is

```
welchADF.test(formula, response, between.s, within.s = NULL, subject = NULL,
     contrast = c("omnibus", "all.pairwise"), effect = NULL,
     correction = c("hochberg", "holm"), trimming = FALSE, per = 0.2,
     bootstrap = FALSE, numsim_b = 999, effect.size = FALSE, numsim_es = 999,
     scaling = TRUE, standardize.effsz = TRUE, alpha = 0.05, seed = 0, ...)
```

We summarize below the meaning of the arguments; the reader may refer to the package documentation for further detail. Note only the three first arguments are required.

A	B	X	W	Subject	Y_1	Y_2
A_1	B_1	X_1	W_1	1	$(1)Y_1^{A_1B_1X_1W_1}$	$(1)Y_2^{A_1B_1X_1W_1}$
⋮	⋮	⋮	⋮	1	⋮	⋮
A_1	B_1	X_1	W_w	1	$(1)Y_1^{A_1B_1X_1W_w}$	$(1)Y_2^{A_1B_1X_1W_w}$
A_1	B_1	X_2	W_1	1	$(1)Y_1^{A_1B_1X_1W_1}$	$(1)Y_2^{A_1B_1X_1W_1}$
⋮	⋮	⋮	⋮	1	⋮	⋮
A_1	B_1	X_x	W_w	1	$(1)Y_1^{A_1B_1X_xW_w}$	$(1)Y_2^{A_1B_1X_xW_w}$
A_1	B_1	X_1	W_1	2	$(2)Y_1^{A_1B_1X_1W_1}$	$(2)Y_2^{A_1B_1X_1W_1}$
⋮	⋮	⋮	⋮	2	⋮	⋮
A_1	B_1	X_x	W_w	2	$(2)Y_1^{A_1B_1X_xW_w}$	$(2)Y_2^{A_1B_1X_xW_w}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮
A_1	B_1	X_x	W_w	$n_{A_1B_1}$	$(n_{A_1B_1})Y_1^{A_1B_1X_xW_w}$	$(n_{A_1B_1})Y_2^{A_1B_1X_xW_w}$
A_1	B_2	X_1	W_1	$n_{A_1B_1}+1$	$(n_{A_1B_1})Y_1^{A_1B_1X_xW_w}$	$(n_{A_1B_1})Y_2^{A_1B_1X_xW_w}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮
A_a	B_b	X_x	W_x	N	$(N)Y_1^{A_aB_bX_xW_w}$	$(N)Y_2^{A_aB_bX_xW_w}$

Table 1: An example dataset with two between-subjects factors A, B, with a and b different levels, respectively; two within-subjects factors X, W with x and w different levels, and a multivariate response with $p = 2$ correlated response columns Y_1, Y_2. The total number of subjects is N.

- `formula` is a data frame object containing the observations and the level combination to which they correspond. The next four arguments refer to the column names.
- `response,between.s,within.s,subject` are strings (or string vectors) indicating the column names for the response, the between-subjects effect(s), the within-subject(s) effects, and the subject column that stores which subject corresponds to each row (hence it cannot be a vector but a single string). In case the design is multivariate, the response will be a vector of columns, one for each response variable. A sample data frame is displayed in Table 1. Each cell A_iB_j has $n_{A_iB_j}$ subjects, and $n_{A_iB_j} \cdot x \cdot w$ rows in the data frame. Here, $N = \sum_{i,j} n_{A_iB_j}$ subjects.
- `contrast` refers to the type of contrast to be performed. Both in `"omnibus"` and `"all.pairwise"` contrasts, the corresponding contrasts matrices are automatically computed as described in Section 2.2.1.
- `effect` is the effect (i.e. column name) involved in the selected contrast. If `effect` is a vector with length 2 or greater and `contrast = "omnibus"`, then an omnibus contrast on an interaction effect will be tested involving simultaneously all the effects of the vector. If `contrast = "all.pairwise"`, then `effect` must have length 1 or 2 to indicate a single effect or a two-way interaction to which tetrad contrasts will be applied; otherwise an error will be thrown. If left blank, the contrast will be applied separately to all of the existing effects and their interactions.
- The rest of arguments specify whether trimmed means and Winsorized variances will be used and the percentage of trimming (`use.robust.estimators, per`), whether bootstrapping should be used to compute an empirical critical value and how many iterations to do (`use.bootstrap, numsim_b`), and whether the effect size and a confidence interval should be computed (again via bootstrap). Effect size allows the choice of using scaling (`scaling = TRUE`) or not (if not, $\eta = 1$ in Eq. 10) and the number of effect-size bootstrap simulations (`effect.size, scaling, numsim_es, loc1, loc2`).

The aforementioned function is an R wrapper that configures the parameters needed for each type of problem by two private functions, which lie at the core of our package. Both are R translations of the SAS functions `wjglm` and `bootcom` and have been named almost the same. Function `wjglm` is used in all cases (including those in which bootstrapping is needed), except when bootstrap is applied to obtain an empirical critical value for a *family* of contrasts. A modification of function `bootcom` is only invoked in that case in order to control family-wise type I error rate (FWER) via percentile bootstrapping. This scenario arises when performing all pairwise contrasts or tetrad contrasts via bootstrap (i.e. `contrast = "all.pairwise",bootstrap = TRUE`).

The prototype of the S3 method for class `formula` is as usual:

```
welchADF.test(formula,data,subset,...)
```

where `data` is a data frame following the same rules as described above for the `formula` parameter of the default method, `subset` is an indexing vector to indicate which rows of `data` should be used

(all by default), and ... stands for the rest of arguments accepted by `welchADF.test.default` and described above to configure the behavior of the test. As with other models, the terms in `formula` are first sought in `data` and then in the environment of the formula. Note, however, that only between-subjects and within-subjects effects and interactions can be specified together with a Subject column when conducting a WJ test, but no model is fit to the data. For this reason, `formula` should be understood only as a way to indicate the factors involved and their nature (between- or within-subjects) but not as a description of a particular model structure. The presence or absence of an interaction in the formula only affects which effects are tested when `contrast = "omnibus",effect = NULL`; otherwise it does not affect at all. The structure should mirror that of the **lme4** package, e.g.

```
welchADF.test(cbind(visits,time,latency) ~ nurs*tunnel + (tunnel|Subject), miceData)
```

means that there is a multivariate response composed of three correlated variables `visits`, `time`, `latency`, and the design has one between-subjects factor `nurs` (because it appears outside but not inside the parenthesis term) and one within-subjects factor `tunnel` because it appears inside the parenthesis. While a within-subjects effect may appear outside the parenthesis to indicate an interaction with a between-subjects effect, between-subjects must not appear inside the parenthesis.

The function returns an object of class `welchADFt`, which is actually a tagged list of lists, one sub-list per effect in an omnibus contrast, or per category involved a pairwise contrast of a given effect. The call is also stored as the last element of the upper-level list with the name `call`, no matter the S3 method employed (be it `welchADF.test.default`, or the ones for class `formula`, `lm`, `aov` or `lmer`). This allows to implement S3 method `update` for class `welchADFt`, no matter which S3 method was called to calculate the model object [6]. Each sub-list has elements named `welch.T`, `numeratorDF`, `denominatorDF`, `contrast.matrix`, `mean.vector`, `sigma.matrix` which store, respectively, the value of the T_{WJ}/c statistic (or $T^{(t)}_{WJ}/c$ if the trimmed version was used), the approximate degrees of freedom of the numerator and denominator, the contrast matrix $\mathbf{R}$ obtained as $\mathbf{R} = \mathbf{C} \otimes \mathbf{U}^T$, and the estimates $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$. It also stores the user arguments when the function was called and, in case the user asked for the effect size, it provides the effect size along with a confidence interval. Refer to the package documentation for further detail.

The package implements S3 methods `summary`, `format` and `print` for objects of class `welchADFt`, as well as other methods widely used on model objects such as `confint` to get confidence intervals of the effect size (in case the user requested to compute it), `model.frame` to extract the input data frame, and `formula` to extract the formula (not available if the object was generated by `welchADF.test.default`).

Case studies

All the datasets analyzed in this section were mentioned in examples designed by the authors of the SAS implementation (Lix and Keselman, 1995), and have also been included in our R package. For that reason it is not necessary to explicitly read them from text files. They are described in detail in the package documentation.

Univariate one-way between-subject design

The dataset was artificially created by Lix et al. and can be downloaded from her personal website[7]. The data recreate those reported by a real study on perception and concentration, on which 42 students were given several puzzles to be solved. The students are divided into three balanced groups as they had previously been asked to imagine solving puzzles in the distant future, near future, or not to imagine anything at all (control group). The response variable represents the number of puzzles each student was able to solve, out of 12. The data are delivered in our package in a variable named `perceptionData`.

This design presents one between-subjects variable, namely the Group to which the student belongs, and no within-subjects variables as each student is measured on only one response variable and then the student is never measured again. The following R commands demonstrate the types of analyses that can be done with this dataset. The results can be checked on the PDF file of the footnote.

```
> str(perceptionData)
'data.frame':   42 obs. of  2 variables:
 \$ Group: Factor w/ 3 levels "control","distantFuture",..: 2 2 2 2 2 2 2 2 2 2 ...
 \$ y    : int  7 5 8 9 8 8 7 7 6 2 ...
> omnibus_LSM <- welchADF.test(perceptionData, response = "y", between.s = "Group")
```

[6] The update should be done in accordance with the function that generated the `welchADFt` object, i.e. passing a formula is not allowed if the object was generated by the default method, and vice-versa.

[7] http://homepage.usask.ca/~lml321/Example1.pdf

```
> summary(omnibus_LSM, verbose = TRUE)
Call:
   welchADF.test(formula = perceptionData, response = "y", between.s = "Group")

Welch-James Approximate DF Test (Least squares means & variances)
Omnibus test(s) of effect and/or interactions

      WJ statistic Numerator DF Denominator DF Pr(>WJ)
Group        1.795            2          24.16  0.1875
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The omnibus contrast on the between-subjects effect Group determined it is not statistically significant when using least-square means. But if we apply trimming:

```
> omnibus_trimmed <- update(omnibus_LSM, trimming = TRUE)
> omnibus_trimmed_boot <- update(omnibus_trimmed, bootstrap = TRUE, seed = 12345)
> summary(omnibus_trimmed)
Call:
   welchADF.test(formula = perceptionData, response = "y", between.s = "Group",
    trimming = TRUE)

      WJ statistic Numerator DF Denominator DF Pr(>WJ)
Group        4.975            2          16.11 0.02076 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

By applying trimmed means and Winsorized variances we do get a statistically significant result. Hence, since the omnibus test was significant on this factor, we do pairwise comparisons on it in order to test which pairs of group levels make the associated groups of responses statistically different. Since the result was obtained with trimming, we continue with it in pairwise comparisons. Only the result of non-bootstrapped trimming is displayed here.

```
> pairwise_trimmed <- welchADF.test(y ~ Group, data = perceptionData, effect = "Group",
  contrast = "all.pairwise", trimming = TRUE, effect.size = TRUE)
> pairwise_trimmed_boot <- update(pairwise_trimmed, bootstrap = TRUE, seed = 12345)
> summary(pairwise_trimmed)
Call:
   welchADF.test(formula = y ~ Group, data = perceptionData, effect = "Group",
    contrast = "all.pairwise", trimming = TRUE, effect.size = TRUE)

                        WJ statistic Numerator DF Denominator DF eff.size adj.pval
control:nearFuture          0.004398            1         10.089 -0.02674   0.9484
distantFuture:nearFuture    0.876366            1          9.778  0.38620   0.7435
control:distantFuture      10.088968            1         17.510 -1.09981   0.0161 *
---
Signif. codes (Hochberg p-values):  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The output shows that the responses associated to the control group significantly differs from those associated to the distantFuture group. The confidence intervals on the effect size can be retrieved as

```
> confint(pairwise_trimmed)
                           2.5 %   97.5 %
control:nearFuture       -0.8208  0.85290
distantFuture:nearFuture -0.3402  1.74365
control:distantFuture    -2.2571 -0.09678
```

An important issue arises in this example that justifies again the use of trimmed estimators. As can be seen in the omnibus tests, the Group is not significant with Least-squares means but it is when we use trimmed means and Winsorized variances. This yields a significant result which could not be detected unless trimming is applied. As the result of the omnibus test with trimmed means is significant, we proceed to the pairwise comparisons using trimming as well. This yields that `control` and `distantFuture` have associated significantly different values of the number of puzzles solved by the students on average.

Two-way factorial (between-subjects) design

Once again, this dataset[8] was artificially created by Lix et al. Quoting from the PDF, the author used summary data presented by Wicherts et al. (2005). These authors examined the effects of stereotype threat on women's mathematics ability. Study participants were assigned to one of six groups defined by crossing the independent factors of test condition (control, nullified, stereotype threat) and sex (male, female). Originally there were four different tests administered to study participants (arithmetic, number series, word problems, and sums tests) the dataset contains only scores for the arithmetic test (out of 40) because these scores exhibited a greater magnitude of variance heterogeneity than scores for the other tests. It is an unbalanced design with cell sizes ranging from 45 to 50 participants, and a total sample size of 283. The data are delivered in our package in a variable named `womenStereotypeData`.

The output of the omnibus tests using robust estimators (trimmed means and Winsorized variances) with and without bootstrapping is shown in first place. Since the interaction between "condition" and "sex" is significant according to trimmed means, the post-hoc pairwise comparisons (tetrad contrasts) are shown using trimmed means with and without bootstrapping. The results match those presented in pages 5 and 6 of the PDF.

```
> omnibus_LSM <- welchADF.test(womenStereotypeData, response = "y", between.s =
  c("condition", "sex"), contrast = "omnibus")
> omnibus_trimmed <- update(omnibus_LSM, trimming = TRUE)
> summary(omnibus_LSM)
Call:
   welchADF.test(formula = womenStereotypeData, response = "y",
    between.s = c("condition", "sex"), contrast = "omnibus")

              WJ statistic Numerator DF Denominator DF Pr(>WJ)
condition            2.151            2          154.7 0.11986
sex                  2.933            1          216.4 0.08824 .
condition:sex        2.521            2          154.7 0.08368 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> summary(omnibus_trimmed)
Call:
   welchADF.test(formula = womenStereotypeData, response = "y",
    between.s = c("condition", "sex"), contrast = "omnibus",
    trimming = TRUE)

              WJ statistic Numerator DF Denominator DF  Pr(>WJ)
condition            5.205            2          93.38 0.007189 **
sex                  5.754            1         130.06 0.017875  *
condition:sex        3.130            2          93.38 0.048347  *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

In this case, the omnibus test is unclear when using least-square means (p-values only slightly greater than the common significance threshold of 0.05 according to the summary of `omnibus_LSM`) but trimming helps make things clearer. This results in both factors and their interaction being statistically significant at a significance level of 0.05 (see the summary of `omnibus_trimmed` above).

Since the omnibus test confirms the significance of all effects, pairwise comparisons should be done on all of them. Due to space constraints, only the two-way `condition:sex` interaction effect was probed (as `effect = c("condition","sex")` in the call to create the `pairwise_LSM` object that was subsequently updated to account for trimming and bootstrapping). Pairwise contrasts on two-way interactions are also known as tetrad contrasts.

```
> pairwise_trimmed <- welchADF.test(y ~ condition*sex, data = womenStereotypeData,
  contrast = "all.pairwise", effect = c("condition", "sex"), trimming = TRUE)
> pairwise_trimmed_boot <- update(pairwise_trimmed, bootstrap = TRUE, seed = 12345)

> summary(pairwise_trimmed_boot, verbose = TRUE)
Call:
   welchADF.test(formula = y ~ condition * sex, data = womenStereotypeData,
```

[8] http://homepage.usask.ca/~lml321/Example2.pdf

```
    contrast = "all.pairwise", effect = c("condition", "sex"),
    trimming = TRUE, bootstrap = TRUE, seed = 12345)

Welch-James Approximate DF Test (Trimmed means [20% trimming] & Winsorized variances)
Multiple tetrad interaction contrasts with respect to condition x sex interaction
using a Bootstrap Critical Value for FWER control
                                WJ statistic Numerator DF Denominator DF significant?
control:stereotype x female:male       0.4662           1          97.49          no
nullified:stereotype x female:male     1.9846           1          79.63          no
control:nullified x female:male        5.7662           1          88.55         yes

Bootstrap critical value: 5.145
```

Pairwise comparisons on the condition:sex interaction with trimmed bootstrapped means reveal only one significant interaction between the pairs of levels `control:nullified` and `female:male`.

Multivariate (mixed) between- × within-subjects design

The problem and the data are described in Keselman et al. (2003). The data represent the reaction times in milliseconds of children with attention-deficit hyperactivity (ADHD) and normal children when they are presented four kinds of inputs: a target alone or an arrow stimuli incongruent, congruent and neutral to the target. According to the authors, the dataset was artificially generated from the summary measures given in the original study by Jonkman et al. (1999), in groups of 20 and 10 children to create an unbalanced design. The data are delivered in our package in two variables named `adhdData` and `adhdData2`.

One-way multivariate vs univariate mixed model This problem can be approached in two different ways: (a) as a one-way multivariate design, which would be the non-parametric equivalent of MANOVA (multivariate ANOVA), or (b) as a univariate mixed model having one between-subjects factor (the student's group) and one within-subjects factor (with four levels, namely the four stimuli measured in every single student). In case we were analysing the data under parametric assumptions, the second option requires sphericity while MANOVA does not (although it needs more data). On the other hand, mixed models are able to capture the covariance structure of the dependent variables and can be generalized to any number of factors. An in-depth discussion on this topic can be found in chapter 9 of Maxwell and Delaney (2004). Keselman et al. (2003) (page 593) insist on using trimmed means and/or bootstrapping with this kind of models in order to overcome deviations from sphericity.

Our package admits both types of analysis. When dealing with a one-way multivariate design, the data must be formated as in Figure 1(a), while a mixed model requires the more systematic format of Figure 1(b) which is valid for an arbitrary amount of factors of both types. As done in their paper, we will analyze this dataset as a mixed model in which the stimuli are an explicit within-subjects factor.

Implicit within-subjects effect The package admits a third way to indicate the within-subjects effects that simplifies its use. It is common to have a dataset with a within-subjects effect expressed in the form of Figure 1(a). In this case, we may want to consider the within-subjects effect underlying the multivariate response. Reshaping this file to match the structure of Figure 1(b) would require some effort by the user. To avoid this, the function allows indicating that the multivariate response is actually an implicit within-subjects effects by including the word `"multivariate"` in the vector of within-subjects column names (if this argument was empty, then we just set the argument `within.s = "multivariate"`). This can be generalized as follows: if we have K within-subjects effects, we can have $K-1$ columns in the data with their explicit names and levels, and leave one effect to be indicated in the multivariate response. In that case, we set `within.s = c("within1", "within2", ..., "within-k-1", "multivariate")` and pass a multivariate response vector argument because the data must have one response column per level of the K-th within-subjects effect. In the code below, the variables with the termination `_multi` show the equivalent calls. Unless we change the model itself (i.e. consider a mixed model or a multivariate one-way model with no within-subjects factor), the results obtained are the same in all types of analyses (omnibus, pairwise, etc), no matter the structure of the input data file. We demonstrate all the possibilities below.

```
> omnibus_LSM_mixed_implicit <- welchADF.test(adhdData, response = c("TargetAlone",
  "Congruent", "Neutral", "Incongruent"), within.s = "multivariate", between.s = "Group",
  contrast = "omnibus")
> omnibus_LSM_multi_oneway <- welchADF.test(cbind(TargetAlone, Congruent, Neutral,
```

Group	TargetAlone	Incongruent	Congruent	Neutral
Normal	568.52	433.80	658.51	711.33
Normal	1034.82	864.79	639.42	815.18
⋮	⋮	⋮	⋮	
ADHD	707.15	872.39	645.83	677.84

(a) As a one-way multivariate model, stored in variable `adhdData`

Group	Stimulus	Subject	Millisec
Normal	TargetAlone	1	568.52
Normal	Incongruent	1	433.80
Normal	Congruent	1	658.51
Normal	Neutral	1	711.33
⋮	⋮	⋮	
ADHD	TargetAlone	30	707.15
ADHD	Incongruent	30	872.39
ADHD	Congruent	30	645.83
ADHD	Neutral	30	677.84

(b) A a mixed model with one between- × one within-subjects factor, as in `adhdData2`

Figure 1: Two alternative ways of arranging the ADHD data input file (*wide* vs *long* format).

```
  Incongruent) ~ Group, data = adhdData)
> omnibus_LSM_mixed <- welchADF.test(adhdData2, response = "Milliseconds",
  between.s = "Group", within.s = "Stimulus", subject = "Subject", contrast = "omnibus")
> omnibus_LSM_mixed_formula <- welchADF.test(Milliseconds ~ Group*Stimulus +
  (Stimulus|Subject), data = adhdData2)
> omnibus_trimmed_formula <- update(omnibus_LSM_mixed_formula, trimming = TRUE)
> omnibus_trimmed_boot <- update(omnibus_trimmed_formula, bootstrap = TRUE, seed = 12345)
```

Above we have demonstrated the possibilities of an omnibus contrast to both arrangements of the data. The first and third models assume a mixed model, where the within-subjects factor is implicit in `omnibus_LSM_mixed_implicit` (using `adhdData`) and explicit in `omnibus_LSM_mixed` (using `adhdData2`). The second model assumes a multivariate one-way model with no within-subjects effects. The model `omnibus_LSM_mixed_formula` assumes an explicit mixed model described by a formula. The formula interface can be fitted only to data in long format like `adhdData2`. Finally, this model is updated to include trimming, and the resulting updated model is updated again to include bootstrapping as well.

```
> summary(omnibus_LSM_mixed_implicit)
Call:
   welchADF.test(formula = adhdData, response = c("TargetAlone",
    "Congruent", "Neutral", "Incongruent"), between.s = "Group",
    within.s = "multivariate", contrast = "omnibus")

                    WJ statistic Numerator DF Denominator DF  Pr(>WJ)
Group                     0.2249            1          24.84 0.639482
multivariate              5.6591            3          21.02 0.005282 **
Group : multivariate      0.5750            3          21.02 0.637759
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> summary(omnibus_LSM_multi_oneway)
Call:
   welchADF.test(formula = cbind(TargetAlone, Congruent, Neutral,
    Incongruent) ~ Group, data = adhdData)

      WJ statistic Numerator DF Denominator DF Pr(>WJ)
Group       0.4227            4          20.52  0.7904
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The results match those presented in Keselman et al. (2003), page 594. The only significant effect is the Stimulus, which acts as the within-subjects factor. When we consider a one-way multivariate design, with no within-subjects factor, then the Group (the between-subjects factor) is deemed not significant. In order perform all pairwise comparisons between the levels of Stimulus, we proceed as follows (only the bootstrap results are shown due to space constraints). The comparison is statistically significant when the value of the WJ statistic is greater than or equal to the bootstrap critical value. In this case, there were two significant differences, namely Incongruent vs TargetAlone, and Incongruent vs Congruent, as mentioned in page 595 of the aforementioned work.

```
> pairwise_trimmed_formula <- update(omnibus_trimmed_formula, contrast = "all.pairwise",
  effect = "Stimulus")
> pairwise_trimmed_formula_boot <- update(pairwise_trimmed_formula, bootstrap = TRUE,
  seed = 123456)

> summary(pairwise_trimmed_formula_boot)
Call:
   welchADF.test(formula = Milliseconds ~ Group * Stimulus + (Stimulus |
    Subject), data = adhdData2, trimming = TRUE, contrast = "all.pairwise",
    effect = "Stimulus", bootstrap = TRUE, seed = 123456)

                       WJ statistic Numerator DF Denominator DF significant?
Congruent:TargetAlone        3.3278            1         15.515           no
Incongruent:TargetAlone     17.3549            1         15.436          yes
Neutral:TargetAlone          0.8251            1          8.419           no
Congruent:Neutral            0.1818            1          8.852           no
Incongruent:Neutral         13.9212            1         15.305          yes
Congruent:Incongruent        8.0013            1         15.503           no

Bootstrap critical value: 8.577
```

Note that, when using the implicit within-subjects effect format, we have to specify `effect = "multivariate"` to indicate that the effect to be tested is the within-subjects effect, even though there is no column with such name in our data. In case we are using the formula interface, this is not available because formula terms must strictly correspond to column names or variables in the environment of the formula.

Multivariate within-subjects design (doubly multivariate)

The three case studies addressed before are probably the most common in practice. Nevertheless, and with the aim of demonstrating the flexibility of the **welchADF** package, we present here an additional, not-so-common setting, namely a doubly multivariate design also proposed by Lix and Keselman (1995). In general, this design arises when several dependent variables are measured for each individual at several time points (every variable measured at each time point), or under different conditions; the latter is the case of the example addressed below.

We could not access the data employed by Lix, hence we have analyzed data from Wuensch (1992) which are freely available in the author's website[9]. In this experiment, wild strain house mice were, at birth, cross fostered onto house mouse (Mus), deer mouse (Peromyscus) or rat (Rattus) nursing mothers. Ten days after weaning, each subject was tested in an apparatus that allowed it to enter four different tunnels: one scented with clean pine shavings, and the other three tunnels with shavings bearing the scent of Mus, Peromyscus, or Rattus respectively. Three variables were measured for each tunnel: the number of visits to the tunnel during a twenty minute test, the time spent by each subject in each of the four tunnels and the latency to first visit of each tunnel.

In this design, the type of nursing mother is a between-subjects factor. The within-subjects factor is scent, with four levels (clean, Mus, Peromyscus, and Rattus). The multivariate response is composed of visits, time, and latency for each tunnel. With this approach, the multivariate response is not treated as another within-subjects factor. The data are delivered in our package in a variable named `miceData`.

```
> head(miceData)
  Subject nurs       tunnel visits   time latency
1       1  Mus        Clean      6 721.35  207.90
2       1  Mus        MusSc      4 318.15   26.78
3       1  Mus PeromyscusSc      2  48.83 1025.33
4       1  Mus     RattusSc      0   0.00 1212.75
5       2  Mus        Clean      8 119.70  685.13
6       2  Mus        MusSc      7 207.90  113.40
```

We first do an omnibus contrast. In the second call we demonstrate how the formula interface can be used in this design to obtain exactly the same result.

```
> omnibus_LSM <- welchADF.test(miceData, response = c("visits", "time", "latency"),
  between.s = "nurs", within.s = "tunnel", subject = "Subject", contrast = "omnibus")
```

[9] http://core.ecu.edu/psyc/wuenschk/SPSS/TUNNEL4b.sav

```
> omnibus_LSM_formula <- welchADF.test(cbind(visits, time, latency) ~ nurs*tunnel +
  (tunnel | Subject), data = miceData)

> summary(omnibus_LSM_formula)
Call:
   welchADF.test(formula = cbind(visits, time, latency) ~ nurs *
    tunnel + (tunnel | Subject), data = miceData)

             WJ statistic Numerator DF Denominator DF   Pr(>WJ)
nurs                4.008            6          21.46 0.0076171  **
tunnel              5.201            9          22.08 0.0007601 ***
nurs : tunnel       5.153           18          21.38 0.0002407 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The least-square means were able to deem all effects statistically significant. Therefore we move to pairwise contrasts for each of the effects, with and without trimming. The pairwise results of the interaction term are not displayed as they consist of a large table.

```
> pairwise_LSM_nurs <- update(omnibus_LSM_formula, effect = "nurs",
  contrast = "all.pairwise")
> pairwise_LSM_tunnel <- update(pairwise_LSM_nurs, effect = "tunnel")
> pairwise_tunnel_trimmed <- update(pairwise_LSM_tunnel, trimming = TRUE)
> pairwise_nurs_trimmed <- update(pairwise_LSM_nurs, trimming = TRUE)

> summary(pairwise_LSM_nurs)
Call:
   welchADF.test(formula = cbind(visits, time, latency) ~ nurs *
    tunnel + (tunnel | Subject), data = miceData, effect = "nurs",
    contrast = "all.pairwise")

                 WJ statistic Numerator DF Denominator DF adj.pval
Mus:Rattus              4.210            3          16.85  0.04287 *
Peromyscus:Rattus       6.141            3          17.34  0.01468 *
Mus:Peromyscus          1.255            3          17.44  0.32030
---
Signif. codes (Hochberg p-values):  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> summary(pairwise_LSM_tunnel)
Call:
   welchADF.test(formula = cbind(visits, time, latency) ~ nurs *
    tunnel + (tunnel | Subject), data = miceData, effect = "tunnel",
    contrast = "all.pairwise")

                     WJ statistic Numerator DF Denominator DF adj.pval
Clean:RattusSc             0.9554            3          24.71  0.42925
MusSc:RattusSc             6.7816            3          23.36  0.01129 *
PeromyscusSc:RattusSc      2.4812            3          24.74  0.33190
Clean:PeromyscusSc         1.2393            3          22.83  0.42925
MusSc:PeromyscusSc         2.2319            3          23.84  0.33190
Clean:MusSc                3.0873            3          24.97  0.22712
---
Signif. codes (Hochberg p-values):  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We will now incorporate bootstrapping to the pairwise comparisons and check the results:

```
> pairwise_nurs_trimmed_boot <- update(pairwise_nurs_trimmed, bootstrap = TRUE, seed = 123)
> pairwise_tunnel_trimmed_boot <- update(pairwise_nurs_trimmed_boot, effect = "tunnel")

> summary(pairwise_nurs_trimmed_boot)
Call:
   welchADF.test(formula = cbind(visits, time, latency) ~ nurs *
    tunnel + (tunnel | Subject), data = miceData, effect = "nurs",
    contrast = "all.pairwise", trimming = TRUE, bootstrap = TRUE,
    seed = 123)
```

```
                 WJ statistic Numerator DF Denominator DF significant?
Mus:Rattus             3.6147            3         10.662           no
Peromyscus:Rattus      6.3409            3          9.551          yes
Mus:Peromyscus         0.6982            3         10.438           no

Bootstrap critical value: 6.292

> summary(pairwise_tunnel_trimmed_boot)
Call:
   welchADF.test(formula = cbind(visits, time, latency) ~ nurs *
    tunnel + (tunnel | Subject), data = miceData, effect = "tunnel",
    contrast = "all.pairwise", trimming = TRUE, bootstrap = TRUE,
    seed = 123)

                     WJ statistic Numerator DF Denominator DF significant?
Clean:RattusSc              1.544            3          15.76           no
MusSc:RattusSc              6.053            3          15.20          yes
PeromyscusSc:RattusSc       3.729            3          15.03           no
Clean:PeromyscusSc          3.106            3          14.26           no
MusSc:PeromyscusSc          1.976            3          15.50           no
Clean:MusSc                 4.842            3          14.12           no

Bootstrap critical value: 5.922
```

The pairwise comparisons using least-squares means are able to detect significant differences only between MusSc and RattusSc in the tunnel effect, and the trimmed-bootstrapped comparison also supports this fact and negates significant differences for any other pair of levels. Regarding the nurs effect, the LSM comparison reveals significant differences in Mus vs Rattus and also in Peromyscus vs Rattus, but the trimmed-bootstrapped only agrees with the latest.

Conclusions and further work

This contribution has demonstrated the applicability of the new **welchADF** package in a variety of experimental designs, ranging from the most simple one, namely a univariate one-way between-subjects design, to a more exotic one like a doubly-multivariate design. The unified approach of Johansen (1980) that has been implemented here leads to a great ease of use for any case study. We have shown in the example code that specifying the factors involved in the design and the type of analysis are done in a straightforward way, and then the code automatically generates the contrast matrices needed and runs the test, no matter how complex the user's design is. Therefore, researchers from other areas may find it more friendly and hence, our effort may contribute to the diffusion of the Welch-James ADF test in applied studies.

In the future, an enhancement may be added so that custom contrasts can be done in addition to the most common omnibus and pairwise contrasts. This requires designing a simple, yet powerful mechanism for the user to describe the desired test in the function arguments.

Acknowledgments

The author wants to thank the editor and two anonymous reviewers for their useful comments and code snippets which have greatly contributed to improve the quality of the package and its integration within the R package ecosystem, and the readability of the manuscript.

Bibliography

Publication Manual of the American Psychological Association, 6th edition, 2013. [p7]

Nlme: Linear and Nonlinear Mixed Effects Models, 2017. URL https://CRAN.R-project.org/package=nlme. R package version 3.1-131. [p3]

J. Algina. Generalization of Improved General Approximation Tests to Split-Plot Designs with Multiple between-Subjects Factors and/or Multiple within-Subjects Factors. *British Journal of*

Mathematical and Statistical Psychology, 50:243–252, 1997. URL https://doi.org/10.1111/j.2044-8317.1997.tb01144.x. [p2, 3]

J. Algina and S. F. Olejnik. Implementing the welch-james procedure with factorial designs. *Educational and Psychological Measurement*, 44(1):39–48, 1984. URL https://doi.org/10.1177/0013164484441004. [p5]

J. M. Aronoff, D. J. Freed, L. M. Fisher, I. Pal, and S. D. Soli. The effect of different cochlear implant microphones on acoustic hearing individuals' binaural benefits for speech perception in noise. *Ear Hear*, 32(4):468–484, 2011. URL https://doi.org/10.1097/AUD.0b013e31820dd3f0. [p3]

D. Bates, M. Mächler, B. Bolker, and S. Walker. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67(1):1–48, 2015. URL https://doi.org/10.18637/jss.v067.i01. [p3]

D. Bates, M. Maechler, B. Bolker, and S. Walker. *lme4: Linear Mixed-Effects Models Using 'Eigen' and S4*, 2016. URL https://CRAN.R-project.org/package=lme4. R package version 1.1-12. [p3]

T. M. Beasley. Multivariate Aligned Rank Test for Interactions in Multiple Group Repeated Measures Designs. *Multivariate Behavioral Research*, 37(2):197 – 226, 2002. URL https://doi.org/10.1207/S15327906MBR3702_02. [p2]

B. M. Bolker, M. E. Brooks, C. J. Clark, S. W. Geange, J. R. Poulsen, M. H. H. Stevens, and J.-S. S. White. Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in Ecology & Evolution*, 24(3):127–135, 2009. URL https://doi.org/10.1016/j.tree.2008.10.008. [p1]

J. Coates and S. McKenzie-Mohr. Out of the Frying Pan, Into the Fire: Trauma in the Lives of Homeless Youth Prior to and During Homelessness. *Journal of Sociology & Social Welfare*, 37(4):65–96, 2010. [p1]

W. J. Conover. The rank transformation - an easy and intuitive way to connect many nonparametric methods to their parametric counterparts for seamless teaching introductory statistics courses. *Wiley Interdisciplinary Reviews: Computational Statistics*, 4(5):432–438, 2012. URL https://doi.org/10.1002/wics.1216. [p2]

J. Dien. The ERP PCA Toolkit: An Open Source Program for Advanced Statistical Analysis of Event-Related Potential Data. *Journal of Neuroscience Methods*, 187(1):138–145, 2010. URL https://doi.org/10.1016/j.jneumeth.2009.12.009. [p3]

J. Dien, M. S. Franklin, C. A. Michelson, L. C. Lemene, C. L. Adams, and K. A. Kiehlf. fMRI Characterization of the Language Formulation Area. *Brain Research*, 1229:179–192, 2008. URL https://doi.org/10.1016/j.brainres.2008.06.107. [p3]

D. M. Erce-Hurn and V. M. Mirosevich. Modern Robust Statistical Methods: An Easy Way to Maximize the Accuracy and Power of Your Research. *American Psychologist*, 63(7):591–601, 2008. URL https://doi.org/10.1037/0003-066X.63.7.591. [p2]

D. A. Fournier, H. J. Skaug, J. Ancheta, J. Ianelli, A. Magnusson, M. N. Maunder, A. Nielsen, and J. Sibert. AD Model Builder: Using automatic differentiation for statistical inference of highly parameterized complex nonlinear models. *Optimization Methods and Software*, 27(2):233–249, 2012. URL https://doi.org/10.1080/10556788.2011.597854. [p3]

J. J. Higgins. *Introduction to Modern Nonparametric Statistics*. Cengage Learning, 2003. [p1]

J. J. Higgins and S. Tashtoush. An aligned rank transform test for interaction. *Nonlinear World*, 1: 201–221, 1994. [p2]

B. H. Huang and S.-A. Jun. Age Matters, and so May Ratters: Rater Differences in the Assessment of Foreign Accents. *Studies in Second Language Acquisition*, 37(4):623 – 650, 2015. URL https://doi.org/10.1017/S0272263114000576. [p3]

H. Huynh. Some approximate test for repeated measurement designs. *Psychometrika*, 43(2):161–175, 1978. URL https://doi.org/10.1007/BF02293860. [p1, 2]

S. Johansen. The Welch-James Approximation to the Distribution of the Residual Sum of Squares in a Weighted Linear Regression. *Biometrika*, 67:85–92, 1980. URL https://doi.org/10.2307/2335320. [p2, 5, 16]

L. M. Jonkman, C. Kemner, M. N. Verbaten, H. van Engeland, J. L. Kenemans, G. Camfferman, J. K. Buitelaar, and H. S. Koelega. Perceptual and response interference in children with attention-deficit hyperactivity disorder, and the effects of methylphenidate. *Psychophysiology*, 36(4):419 – 429, 1999. URL https://doi.org/10.1111/1469-8986.3640419. [p12]

J. Kayser, C. E. Tenke, C. J. Kroppmann, D. M. Alschuler, S. Fekri, S. Ben-David, C. M. Corcoran, and G. E. Bruder. Auditory event-related potentials and α oscillations in the psychosis prodrome: Neuronal generator patterns during a novelty oddball task. *International Journal of Psychophysiology*, 91(2):104–120, 2014. URL https://doi.org/10.1016/j.ijpsycho.2013.12.003. [p3]

H. J. Keselman, R. K. Kowalchuk, J. Algina, L. M. Lix, and R. R. Wilcox. Testing treatment effects in repeated measures designs: Trimmed means and bootstrapping. *British Journal of Mathematical and Statistical Psychology*, 53(2):175–191, 2000. URL https://doi.org/10.1348/000711000159286. [p4, 7]

H. J. Keselman, R. R. Wilcox, and L. M. Lix. A generally robust approach to hypothesis testing in independent and correlated groups designs. *Psychophyisiology*, 40:586–596, 2003. URL https://doi.org/10.1111/1469-8986.00060. [p2, 3, 4, 7, 12, 13]

H. J. Keselman, J. Algina, L. M. Lix, R. R. Wilcox, and K. N. Deering. A generally robust approach for testing hypotheses and setting confidence intervals for effect sizes. *Psychological Methods*, 13(2): 110–129, 2008. URL https://doi.org/10.1037/1082-989X.13.2.110. [p3, 7]

M. Koller. *Robust Estimation of Linear Mixed Models*. PhD thesis, ETH Zurich, 2013. URL http://e-collection.library.ethz.ch/eserv/eth:6670/eth-6670-02.pdf. [p3]

M. Koller. robustlmm: An R Package for Robust Estimation of Linear Mixed-Effects Models. *Journal of Statistical Software*, 75(6):1–24, 2016. URL https://doi.org/10.18637/jss.v075.i06. [p3]

M. Koller. *robustlmm: Robust Linear Mixed Effects Models*, 2017. URL https://CRAN.R-project.org/package=robustlmm. R package version 2.1-3. [p3]

J. R. Levin. Overcoming feelings of powerlessness in "aging" researchers: a primer on statistical power in analysis of variance designs. *Psychology and aging*, 12(1):84–106, 1997. URL https://doi.org/10.1037/0882-7974.12.1.84. [p1]

L. M. Lix and H. J. Keselman. Approximate Degrees of Freedom Tests: A Unified Perspective on Testing for Mean Equality. *Psychological Bulletin*, 117(3):547–560, 1995. URL https://doi.org/10.1037/0033-2909.117.3.547. [p1, 2, 3, 4, 9, 14]

M. Maechler, P. Rousseeuw, C. Croux, V. Todorov, A. Ruckstuhl, M. Salibian-Barrera, T. Verbeke, M. Koller, E. L. T. Conceicao, and M. Anna di Palma. *robustbase: Basic Robust Statistics*, 2016. URL http://robustbase.r-forge.r-project.org/. R package version 0.92-7. [p2]

P. Mair and R. Wilcox. *WRS2: A Collection of Robust Statistical Methods*, 2017. URL https://CRAN.R-project.org/package=WRS2. R package version 0.9-2. [p2]

R. A. Maronna, D. R. Martin, and V. J. Yohai. *Robust Statistics: Theory and Methods*. John Wiley & Sons, 2006. [p2]

S. E. Maxwell and H. D. Delaney. *Designing Experiments and Analyzing Data: A Model Comparison Perspective, 2nd Ed.* Routledge, 2004. [p12]

G. W. Milligan, D. S. Wong, and P. A. Thompson. Robustness properties of nonorthogonal analysis of variance. *Psychological Bulletin*, 101(3):464–470, 1987. URL https://doi.org/10.1037/0033-2909.101.3.464. [p1]

U. C. Müller, P. Asherson, T. Banaschewski, J. K. Buitelaar, R. P. Ebstein, J. Eisenberg, M. Gill, I. Manor, A. Miranda, R. D. Oades, H. Roeyers, A. Rothenberger, J. A. Sergeant, E. J. Sonuga-Barke, M. Thompson, S. V. Faraone, and H.-C. Steinhausen. The Impact of Study Design and Diagnostic Approach in a Large Multi-Centre ADHD Study. Part 1: ADHD Symptom Patterns. *BMC Psychiatry*, 11(54), 2011. URL https://doi.org/10.1186/1471-244X-11-54. [p3]

J. Ruscio and B. Roche. Variance heterogeneity in published psychological research: a review and a new index. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, 8 (1):1–11, 2012. URL https://doi.org/10.1027/1614-2241/a000034. [p2]

M. J. V. Ryzin, E. A. Carlson, and L. A. Sroufe. Attachment discontinuity in a high-risk sample. *Attachment & Human Development*, 13(4):381–401, 2011. URL https://doi.org/10.1080/14616734.2011.584403. [p3]

M. I. Salazar-Alvarez, V. G. Tercero-Gómez, M. del Carmen Temblador-Pérez, A. E. Cordero-Franco, and W. J. Conover. Nonparametric analysis of interactions: a review and gap analysis. In *Proceedings of the 2014 Industrial and Systems Engineering Research Conference*, 2014. [p2]

M. Salibián-Barrera, S. Van Aelst, and G. Willems. Fast and robust bootstrap. *Statistical Methods and Applications*, 17(1):41–71, 2008. URL https://doi.org/10.1007/s10260-007-0048-6. [p4]

SAS Institute. *SAS/STAT 9.3 User's Guide*, 2011. [p3]

S. S. Sawilowsky. Nonparametric Tests of Interaction in Experimental Design. *Review of Educational Research*, 60(1):91–126, 1990. URL https://doi.org/10.3102/00346543060001091. [p2]

K. Singh. Breakdown theory for bootstrap quantiles. *The Annals of Statistics*, 26(5):1719–1732, 1998. URL https://doi.org/10.1214/aos/1024691354. [p4]

H. Skaug, D. Fournier, A. Nielsen, A. Magnusson, and B. Bolker. *glmmADMB: Generalized Linear Mixed Models Using 'AD Model Builder'*, 2016. http://glmmadmb.r-forge.r-project.org, http://admb-project.org. [p3]

L. Symes, J. McFarlane, L. Frazier, M. C. Henderson-Everhardus, G. McGlory, K. B. Watson, Y. Liu, C. E. Rhodes, and R. C. Hoogeveen. Exploring violence against women and adverse health outcomes in middle age to promote women's health. *Critical Care Nursing Quarterly*, 33(3):233–243, 2010. URL https://doi.org/10.1097/CNQ.0b013e3181e6d7c4. [p3]

G. Vallejo and P. Livacic-Rojas. Comparison of Two Procedures for Analyzing Small Sets of Repeated Measures Data. *Multivariate Behavioral Research*, 40(2):179–205, 2005. URL https://doi.org/10.1207/s15327906mbr4002_2. [p2]

G. Vallejo, A. Fidalgo, and P. Fernández. Effects of covariance heterogeneity on three procedures for analyzing multivariate repeated measures designs. *Multivariate Behavioral Research*, 36(1):1–27, 2001. URL https://doi.org/10.1207/S15327906MBR3601_01. [p2]

G. Vallejo, J. Morisa, and N. M. Conejo. A SAS/IML Program for Implementing the Modified Brown-Forsythe Procedure in Repeated Measures Designs. *Computer Methods and Programs in Biomedicine*, 83(3):169–177, 2006. URL https://doi.org/10.1016/j.cmpb.2006.06.006. [p2, 3]

G. Vallejo, M. Ato, M. P. Fernández, and L.-R. P. E. A Practical Method for Analyzing Factorial Designs with Heteroscedastic Data. *Psychological Reports*, 102(3):643–656, 2008. URL https://doi.org/10.2466/pr0.102.3.643-656. [p2]

P. J. Villacorta. *ART: Aligned Rank Transform for Nonparametric Factorial Analysis*, 2015. URL https://CRAN.R-project.org/package=ART. R package version 1.0. [p2]

P. J. Villacorta. *welchADF: Welch-James Statistic for Robust Hypothesis Testing under Heterocedasticity and Non-Normality*, 2017. URL https://CRAN.R-project.org/package=welchADF. R package version 0.2. [p3]

J. Wang, R. Zamar, A. Marazzi, V. Yohai, M. Salibian-Barrera, R. Maronna, E. Zivot, D. Rocke, D. Martin, M. Maechler, and K. Konis. *robust: Port of the S+ "Robust Library"*, 2017. URL https://CRAN.R-project.org/package=robust. R package version 0.4-18. [p2]

B. L. Welch. On the comparison of several mean values: An alternative approach. *Biometrika*, 38: 330–336, 1951. URL https://doi.org/10.2307/2332579. [p2]

J. M. Wicherts, C. V. Dolan, and D. J. Hessen. Stereotype threat and group differences in test performance: a question of measurement invariance. *Journal of Personality and Social Psychology*, 89(5): 696–716, 2005. URL https://doi.org/10.1037/0022-3514.89.5.696. [p11]

R. Wilcox. *Introduction to Robust Estimation & Hypothesis Testing, 3rd Ed.* Academic Press, 2012. [p2]

R. R. Wilcox, H. J. Keselman, and R. K. Kowalchuk. Can tests for treatment group equality be improved?: The bootstrap and trimmed means conjecture. *British Journal of Mathematical and Statistical Psychology*, 51:123–134, 1998. URL https://doi.org/10.1111/j.2044-8317.1998.tb00670.x. [p3]

S. Wood. *Mgcv: Mixed GAM Computation Vehicle with Automatic Smoothness Estimation*, 2017. URL https://CRAN.R-project.org/package=mgcv. R package version 1.8-19. [p4]

S. Wood and F. Scheipl. *Gamm4: Generalized Additive Mixed Models Using 'mgcv' and 'lme4'*, 2017. URL https://CRAN.R-project.org/package=gamm4. R package version 0.2-5. [p4]

K. L. Wuensch. Fostering house mice onto rats and deer mice: Effects on response to species odors. *Animal Learning & Behavior*, 20(3):253 – 258, 1992. URL https://doi.org/10.3758/BF03213379. [p14]

Pablo J. Villacorta
Department of Computer Science and Artificial Intelligence, University of Granada
ETSIIT, C/Periodista Daniel Saucedo Aranda s/n, 18071 Granada, Spain
pjvi@decsai.ugr.es