Checking the Equality of Covariance Matrices: Some Practical Aspects

Version 2.0
(4.12.2020)

Haiko Lüpsen

Regionales Rechenzentrum (RRZK)

Kontakt: Luepsen@Uni-Koeln.de

Universität zu Köln

Checking the Equality of Covariance Matrices: some practical aspects

Abstract

A Monte Carlo simulation has been performed for split-plot designs to study the requirement of equal covariance matrices for ANOVA procedures. First those situations are identified in which the parametric F test and the multivariate tests react sensitive to heterogeneity of the covariance matrices, then reliable tests for homogeneity of covariance matrices are surveyed. The results revealed a remarkable effect of the relation between sample sizes and within group correlations. Therefore in many situations additional tests for homogeneity of correlation matrices are needed. Among the tests considered are Box's M test, three tests by Schott, a Levene-like test and a pure multivariate dispersion test (RD). For the test of correlation matrices the methods by Jennrich and Larntz & Perlman (LP) have been chosen. The Levene-like test, the RD and the LP test showed an attractive performance.

1. Introduction

The analysis of variance (ANOVA) is one of the most important und frequently used methods of applied statistics. Especially split-plot designs, sometimes also called mixed designs, repeated measures designs involving two or more independent groups, are among the most common experimental designs in educational, psychological, medical and many other fields of scientific research. But in contrast to the between subject ANOVA the assumptions are more substantial for mixed designs. These are, besides the independence of the observed units, multivariate normality of the residuals, sphericity (homogeneity of the variances for the repeated measurement factors) and homogeneity of the covariance matrices. In particular the last requirement is essential for the univariate as well as for the multivariate approach, as e.g. Algina (1994) and Keselman et al. (1993) pointed out. Their simulation studies showed that a violation of this assumption may not be severe in balanced designs, however the reaction of the ANOVA tests, especially for the interaction effect, may be dramatic in unbalanced designs. The empirical type I error rises up to 0.10 and higher for a nominal α=0.05, and on the other side the power may be sincerely dampened. And these effects increase if the underlying distribution is nonnormal, and particularly for the multivariate tests. However, they restricted their analyses only to the case that the covariance matrices were multiples of each other, leading essentially to different variances for the groups. But the covariance matrix is a function of the correlation matrix and the vector of the variances of the dependent variables. Therefore heterogeneous covariance matrices may have equal variances but heterogeneous correlations, e.g. constant correlations within each group which differ between the groups, or correlations which vary for each pair of variables as well as between the groups. There are only few studies in regard of the latter type of heterogeneity, to mention here Beasley & Sheehan (1994) which is restricted to equal sample sizes, and Fouladi & Yockey (2002) who analyzed only the 2-group case. But both studies investigated only the impact on multivariate tests and found no remarkable effect of unequal correlation matrices.

Own simulations within this study (see section 3) confirmed the results from Algina (1994) and Keselman et al. (1993). But also the reaction of the parametric F test and the multivariate test on unequal correlations were analyzed, where several situations have to be distinguished. Here, in case of unbalanced designs, attention is paid on the relation between the sample sizes n_i and the correlations $r^{(i)}$ within group i, which results into situations similar to the conditions of positive and negative pairing in designs for the relation between the sample sizes n_i and the vari-

ances s_i^2 [1]. In order to detect unequal correlations $r^{(i)}$ special tests for heterogeneity of correlation matrices are helpful. Therefore both, tests for homogeneity of the covariance matrices as well as tests for homogeneity of the correlation matrices, will be considered in the study.

Unfortunately the alternatives to the parametric analysis of ANOVA designs are very rare in case of inequality of the covariance matrices. There are a few to mention:

- *Welch & James* ANOVA for heterogeneous variances, with the restriction that the test of the interaction is only valid for large sample sizes (see e.g. Keselman et al., 1993, and Keselman et al., 2000), which confines this test, because it is just the interaction effect which is primarily affected by heterogeneous covariance matrices,
- *general approximate procedure* (GA) as well as *improved general approximate procedure (*IGA*)* (e.g. Huynh, 1978 and Algina, 1994), improvements of the Huynh-Feldt degrees of freedom adjustment for the case of unequal covariance matrices,
- generalization of the well-known test for unequal variances by *Brown & Forsythe* to the case of multivariate data (e.g. Algina, 1994),

but corresponding functions in the software packages are rare. Only SAS seems to offer macros for these techniques, and the author offers a R function for the Welch-James procedure. But even if these methods cannot be applied, the user is not released from checking the assumption. For, only if the homogeneity is confirmed, he can trust in the results from the ANOVA. But one has to realize, as will be shown, that the impact of the distribution of the dependent variable is much larger on the homogeneity tests than on the ANOVA tests. The aim of this study is to find a viable way to come to a reliable result for the test of the homogeneity of covariance matrices as a check of the ANOVA assumption.

2. The methods to be compared

At first the model for the mixed design will be given. For one grouping factor A and one repeated measures factor B, often called trial factor, the 2-factorial ANOVA model for a dependent variable *y* shall be denoted by

$$y_{ikj} = \alpha_i + \beta_j + (\alpha\beta)_{ij} + \tau_{ik} + (\beta\tau)_{ikj} + e_{ikj} \qquad (1)$$

with fixed effects α_i (factor A, i=1,..,I), β_j (factor B, j=1,..,J), $(\alpha\beta)_{ij}$ (interaction AB), n_i subjects per group (i=1,..,I), a subject specific variation τ_{ik} (k=1,..,n_i), multivariate normally distributed error e_{ikj} with equal variances and covariance matrices $\Sigma^{(i)}$ of dimension JxJ and $N = \sum n_i$. The assumption to be checked is

$$\Sigma^{(1)} = ... = \Sigma^{(I)}$$

though for the analysis of variance only the following assumption is de facto required

$$C\Sigma^{(1)}C' = ... = C\Sigma^{(I)}C'$$

as Boik (1981) pointed out. Here C is a Jx(J-1) orthonormal contrast matrix. It is obvious that the first hypothesis implies the second one. In this simulation study the tests will be performed for both hypotheses, as the second one is less stringent.

1. The pairing problem: the parametric F-test tends to be conservative, if cells with larger n_i have also larger variances s_i^2 (positive or direct pairing), and reacts liberal, if cells with larger n_i have the smaller variances (negative or inverse pairing), see e.g. Feir & Toothaker (1974)

Tests for the homogeneity of covariance matrices

The following tests are considered, where S_i is the covariance matrix of group i (i=1,..,I) and S their weighted mean:

- *likelihood ratio* (LR) test:

  $$L = (N-I)\log|S| - \sum_i^I (n_i - 1)\log(|S_i|)$$

  L is approximately χ^2 distributed with $J(J+1)(I-1)/2$ df. De facto this is the so-called modified likelihood test, because the factors N and n_i are replaced by (N-1) and (n_i -1). This test is based on the multivariate normal distribution.

- *Box M* test:
  This test is just a modification of the LR test by multiplying L with a term based on I, J and n_i resulting in a statistic denoted with M. On the other hand, this test is a generalization of the Bartlett test for homogeneity of variances to the case of multivariate data. But to test M Box uses the χ^2 distribution only if $\min(n_i)>20$, $I<6$ and $J<6$. Otherwise he approximates M by the F distribution. For details see Box (1949) or, in an easy to read version, Zaiontz.

- *Schott's T_1, T_2, T_3* tests:
  Schott developed several tests for the comparison of covariance matrices. T_1 is based on the multivariate normal distribution, too, comparatively simple and replaces the determinants (which can be computed as the product of the eigenvalues) in the classical LR test by the trace (sum of eigenvalues) of the covariance matrices, which allows the computation also for smaller samples. Nagao (1973) computed a more exact approximation of the test statistic. T_2 and T_3 are based on the elliptical distribution, which includes the normal distribution as a special case, and on the kurtosis to consider the deviation from the normal distribution. While T_2 works with a common value of the kurtosis for all J variables, T_3 allows individual values for each variable. Therefore T_2 and T_3 cover a larger range of distributions than T_1 and the LR tests. For formulae and details see Schott (2001) or Hallin & Paindaveine (2009).

- *Levene* test:
  this is a Levene-like test proposed by O'Brien (1992).

  Let $\tilde{m}_j^{(i)}$ be the median of variable j in group i. For each subject k in group i the following cross products $s_{j_1 j_2}$ (j_1, j_2 =1,...,J) are computed:

  $$s_{k j_1 j_2}^{(i)} = (y_{i k j_1} - \tilde{m}_{j_1}^{(i)})(y_{i k j_2} - \tilde{m}_{j_2}^{(i)})$$

  which are transformed into

  $$\hat{s}_{k j_1 j_2}^{(i)} = \mathrm{sgn}(s_{k j_1 j_2}^{(i)}) \cdot \sqrt{\left|s_{k j_1 j_2}^{(i)}\right|}$$

  where sgn denotes the sign function. Next, for each subject k the lower triangular matrix of $\hat{s}_{k j_1 j_2}^{(i)}$ is transformed into a vector $w_{kj}^{(i)}$ giving a data matrix W with N rows and $(J+1)J/2$ columns. Finally a multivariate analysis of variance is applied on W, e.g. Wilks Lambda test, resulting in a test of the homogeneity of the covariance matrices for groups i=1,...,I.

  Additionally the nonparametric multivariate analysis of variance procedure by Agresti & Pendergast (1986) has been tried.

- *Robust Dispersion* (RD) test:

  this test, also proposed by O'Brien (1992), compares only the variances from each group. But if the test indicates a signficant difference, also the unequality of the covariance matrices can be concluded.

  From the $\hat{s}^{(i)}_{k j_1 j_2}$ as computed above only the variances $\hat{s}^{(i)}_{kjj}$ are used to compute the following scores:

  $$v_{ik} = \sqrt{\sum_j^J s^{2^{(i)}}_{kjj}}$$

  giving a data vector of size *N*. Finally a 1-factorial analysis of variance is applied on these v_{ik} to compare the *J* variances of each group.

Tests for the homogeneity of correlation matrices

The following tests are considered

- *Jennrich* test:
  this is probably the most cited method to test the equality of correlation matrices. The test is based on the multivariate normal distribution and requires large samples, while for small n_i the results are rather poor (Larntz & Perlman, 1985). For the formula see Jennrich (1970).
- *Levene-like* test for equal correlations:
  this Levene-like test, a variant of the Levene test presented above, checks the homogeneity of the correlation matrices R_i for groups *i* (*i*=1,...,*I*).

  Let z_{ikj} be the z-scores of y_{ikj} based on the means and standard deviations of group *i*. Then the following cross products $s_{j_1 j_2}$ (j_1, j_2 =1,...,*J*) are computed:

  $$s^{(i)}_{k j_1 j_2} = z_{ikj_1} z_{ikj_2}$$

  which are transformed into

  $$\hat{s}^{(i)}_{k j_1 j_2} = \operatorname{sgn}(s^{(i)}_{k j_1 j_2}) \cdot \sqrt{\left| s^{(i)}_{k j_1 j_2} \right|}$$

  where sgn denotes the sign function. Next, for each subject *k* the lower triangular matrix of $\hat{s}^{(i)}_{k j_1 j_2}$, without the diagonal values, is transformed into a vector $w^{(i)}_{kj}$ giving a data matrix *W* with *N* rows and (*J*-1)*J*/2 columns. Finally a multivariate analysis of variance is applied on *W*, e.g. Wilks Lambda test, resulting in a test of the homogeneity of the correlation matrices for groups *i*=1,...,*I*.
- *Larntz & Perlman* test (*LP*):
  in contrary to the Jennrich test this one is conceived for small samples. For details see Larntz & Perlman (1985).

  Let $r^{(i)}_{j_1 j_2}$ be the correlation of variables j_1 and j_2 in group *i*, $z^{(i)}_{j_1 j_2} = \frac{1}{2}\log\left(\frac{1 + r^{(i)}_{j_1 j_2}}{1 - r^{(i)}_{j_1 j_2}}\right)$

  the z transformed correlation, and for the upper off-diagonal matrix ($j_1 < j_2$)

$$S_{j_1 j_2} = \sum_i^I (n_i - 3)(z^{(i)}{}_{j_1 j_2})^2 - \frac{\left(\sum_i^I (n_i - 3) z^i{}_{j_1 j_2}\right)^2}{\sum_i^I (n_i - 3)}$$

then the test statistic is $T=\max(S_{j_1 j_2})$, which is approximately χ^2 distributed with I-1 df, but instead of $p=1-\chi^2(T)$ the resulting p value is computed as $p=1-(\chi^2(T))^{(J-1)J/2}$.

- Box M test:
  while Box's test is primarily to check the homogeneity of covariance matrices, it can also be used to check the equality of correlation matrices by transforming the dependent variable y into z scores separately for each group. Thus the covariances become correlations. Accordingly the degrees of freedom for the χ^2 test have to be adjusted to $(J\text{-}1)J/2$, because only the off diagonal values are used.

Remarks

- Due to the computational procedure the LR and the Box tests require minimum cell counts n_i in relation to the number of variables J: $\min(n_i) > J+1$. This makes these two tests inapplicable for very small n_i .
- As the Levene and the robust dispersion tests are based on the raw data, a comparison of the covariance matrices, which are transformed by an orthonormal contrast matrix, is not possible.
- Hallin & Paindaveine (2009) developed three tests, which are equivalent to the three tests T_1, T_2, T_3 by Schott: one based on the multivariate normal distribution and two based on the elliptical distribution, from which one assumes a common kurtosis for all J variables and one individual kurtoses. There are also a couple of tests by Marden & Gao (2002) which are based on spatial signs and ranks. They are not part of this simulation study.

Comparisons in other studies

- Greenstreet & Connor (1974) compared the LR with the Box test. In their simulation study Box's test scored because of its better type I error control for smaller sample sizes, while for larger sample sizes ($N \geq 100$) for both methods the observed error levels come very close to the nominal level. To detect differences Box's test needs considerable sample sizes or effect sizes, while the number of groups or variables has no impact.
- O'Brien (1992) compared the likelihood ratio, the Levene and the robust dispersion test, but only for bivariate distributions. He confirms the poor performance of the LR test for even modest departures from normality, whereas the other two tests provide accurate control over the size for modest departures, but are conservative in the presence of outliers. In regard of the power the best performance is shown by the LR method only for underlying normal distributions, while the Levene and the RD tests are by far better for nonnormal distributions.
- Schott (2001) compared his tests T_1, T_2 and T_3 but only for equal sample sizes and with emphasis upon type I error control. For normal distributions T_1 controlled the error much better than T_2 and T_3 which behaved too liberal for small n_i. In case of elliptical distributions of course T_2 and T_3 were leading, while T_1 had a poor power, and for distributions with a large kurtosis T_3 performed the best.
- Pervaiz & Skinner (1990) evaluated tests based on elliptical distributions and found that they all have a poor type I error control, for underlying normal and even elliptical distributions, in particular for small n_i.

- Larntz & Perlman (1985) compared their own test with three versions of Jennrich's method. They found that their own test is able to keep the nominal α level at least for $n_i \leq 40$, while all versions of Jennrich's test need larger n_i so that the actual error level comes close to the nominal α level. In regard to the power they found no differences between the methods. Unfortunately they give no information about the distributions used in their Monte Carlo study.

3. Design of the study

The parameters of this study are the size of the design (number of cells), cell frequencies (equal, unequal), cell counts (5,10,...,50), the covariance and correlation structure (equal or unequal correlations) and the underlying multivariate distribution. The resulting sample sizes N vary from 15 to 1200. For each situation there are 2000 replications.

Four designs are analyzed:

- a 3*3 design ("small design") with equal cell counts (balanced), and one with unequal cell counts having a ratio $\max(n_i)/\min(n_i)$ of 3.5 (unbalanced), and
- a 4*6 design ("large design") with equal cell counts (balanced), and one with unequal cell counts having a ratio $\max(n_i)/\min(n_i)$ of 4 (unbalanced),

7 different models of multivariate distributions with covariance matrices $\Sigma^{(i)}$ (i=1,..,I) have been chosen. $s_{j_1 j_2}^{(i)}$ denotes the covariances of $\Sigma^{(i)}$, $s_{jj}^{(i)}$ the variances, $r_{j_1 j_2}^{(i)}$ the correlations and $r^{(i)}$ a constant correlation within group i (j, j_1, j_2=1,...,J and group i=1,...I). The following two correlation structures are used:

- exchangeable (equal covariances, compound symmetry) with $r^{(i)}$=0.3, a value that seems realistic and had often been chosen (see e.g. Emrich & Piedmonte, 1992), and
- descending correlations $r_{12}^{(i)},...,r_{56}^{(i)}$=(0.7, 0.5, 0.4, 0.2, 0.1) for large designs, respectively $r_{12}^{(i)}$, $r_{23}^{(i)}$=(0.7, 0.4) for small designs, which is similar to the AR(1) structure (unequal covariances, no sphericity or compound symmetry).

The following distributions have been selected:

- multivariate normal with equal variances $s_{jj}^{(i)}$ (i=1,..,I),
- multivariate exponential (λ=0.4) with μ=2.5, which is highly skewed (skewness=2),
- multivariate uniform in the interval [0,5] rounded to integer values 1,2,...,5,
- multivariate lognormal (μ=0 and σ=0.25) which is slightly skewed (skewness=0.778),
- mixed left and right skewed (transformation $\log_2(1+x)$ with multivariate uniform x in the interval (0,1), where for two levels of A the values have been mirrored at the mean).

To analyze on one side the impact of heterogeneous covariance matrices $\Sigma^{(i)}$ (i=1,..,I) on the behavior of the ANOVA methods, and on the other side to compare the power of the homogeneity tests listed above, several models of heterogeneous covariance matrices $\Sigma^{(i)}$ (i=1,..,I) have been chosen. Here the sample sizes n_i play an important role. For unbalanced large designs these are n_i= (6, 4, 2, 8), and for small designs n_i= (7, 2, 6). Here the selected parameters are only listed for large designs. The models are defined as follows:

1) $\Sigma^{(i)} = c_i\Sigma^{(1)}$ and therefore $s_{jj}^{(i)} = c_i\, s_{jj}^{(1)}$ with different sets of coefficients c_i (i=1,..,I) :
   a) c_i = (1.0, 0.75, 0.60, 0.50) (corr(n_i , $s_{jj}^{(i)}$) ~ 0.0), n_i and $s_{jj}^{(i)}$ independent,
   b) c_i = (1.0, 0.75, 0.75, 0.50) (corr(n_i , $s_{jj}^{(i)}$) ~ -0.3), decent negative pairing,
   c) c_i = (1.0, 1.3, 2.0, 1.3) (corr(n_i , $s_{jj}^{(i)}$) ~ -0.7), strong negative pairing,
   d) c_i = (1.0, 1.2, 0.77, 1.5) (corr(n_i , $s_{jj}^{(i)}$) ~ 0.8), strong positive pairing.

2) Unequal within group correlations $r^{(i)}$ for i=1,..,I
   (equal correlations $r_{j_1j_2}^{(i)}$ within each group for j_1, j_2 =1,...,J with $j_1 < j_2$),
   a) $r^{(i)}$= (0.2 , 0.3 , 0.4 , 0.5) (corr(n_i , $r^{(i)}$) ~ 0.1), no pairing,
   b) $r^{(i)}$= (0.6 , 0.15, 0.15, 0.6) (corr(n_i , $r^{(i)}$) ~ 0.9), positive pairing,
   c) $r^{(i)}$= (0.15, 0.6 , 0.6 , 0.15) (corr)n_i , $r^{(i)}$) ~ -0.9), negative pairing.

3) Unequal within group correlations $r^{(i)}$ for i=1,..,I and varying within group correlations $r_{j_1j_2}^{(i)}$ for variables j_1, j_2 (j_1, j_2 =1,...,J with $j_1 < j_2$).

4) combination of models 1 and 2
   a) $\Sigma^{(i)} = c_i\Sigma^{(1)}$ with c_i = (1.0, 1.3, 2.0, 1.3) and corr ~ - 0.7 combined with $r^{(i)}$= (0.15, 0.6, 0.6, 0.15)
   (negative pairing of variances combined with negative pairing of group-correlations)
   b) $\Sigma^{(i)} = c_i\Sigma^{(1)}$ with c_i = (1.0, 1.3, 2.0, 1.3) and corr ~ - 0.7 combined with $r^{(i)}$= (0.6 ,0.15, 0.15, 0.6)
   (negative pairing of variances combined with positive pairing of group-correlations)
   c) $\Sigma^{(i)} = c_i\Sigma^{(1)}$ with c_i = (1.0, 1.2, 0.77, 1.5) and corr ~ +0.8 combined with $r^{(i)}$= (0.15, 0.6, 0.6, 0.15)
   (positive pairing of variances combined with negative pairing of group-correlations)
   d) $\Sigma^{(i)} = c_i\Sigma^{(1)}$ with c_i = (1.0, 1.2, 0.77, 1.5) and corr ~ +0.8 combined with $r^{(i)}$= (0.15, 0.6, 0.6, 0.15)
   (positive pairing of variances combined with negative pairing of group-correlations)

(The above listed values for the coefficients and correlations refer to large designs. Similar values have been chosen for small designs.)

Concerning the impact of the c_i in $\Sigma^{(i)} = c_i\Sigma^{(1)}$ - the c_i play the roll of variances - it has to be reminded that, as studied in the case of between subject designs, mainly the variation of the group variances is responsible for the displeasing effect on the type I error rates of the parametric F test: the larger the variation the larger the disturbance on the F test (see Box, 1954). Preliminary studies showed that the group correlations $r^{(i)}$ behave in a similar way: the larger the variation of the $r^{(i)}$ the larger the impact on the parametric F test and the multivariate tests. Therefore the c_i and $r^{(i)}$ have been chosen in a way that their variation is moderate and typical for real data.

First the type I error rates of the ANOVA procedures are checked for the tests of factor A, B and the interaction AB, applying the following methods: the F test with and without the Huynh-Feldt adjustment, the multivariate tests by Hotelling-Lawley and by Pillai, and for comparing purposes also three nonparametric ANOVA methods, (1) a generalization of the Kruskal-Wallis and the Friedman test (KWF), as proposed by Luepsen (2020), (2) a generalization of the van der Waerden test, as proposed by Luepsen (2020), (3) Koch‘s method for the analysis of split-plot designs (Koch, 1969), a nonparametric version of the multivariate Hotelling test, without a sphericity assumption. Then the type I error rates of the methods to check the homogeneity of covariance matrices are controlled for the two correlation structures. Their power is estimated for all models listed above. The type I error rates and the power of the methods to check the homogeneity of correlation matrices are similarly computed, but only for models 2 and 3. Con-

cerning model 3, with correlations varying for any two variables within each group and varying over groups, a large number of different correlation patterns would be possible. Four different matrices have been chosen, each spreading correlations $0.1 \leq r_{j_1 j_2} \leq 0.5$ within each group. They all led to similar results. Therefore only the findings of one of them are reproduced here.

4. Results

Criteria

A deviation of 25 percent ($\alpha + 0.25\alpha$) - that is 6.25 percent for α=0.05 - can be regarded as a moderate definition of robustness (see Peterson, 2002), whereas 50 percent ($\alpha + 0.5\alpha$) - that is 7.50 percent for α=0.05 - will be treated as liberal robustness, according to Bradleys liberal criterion (see Bradley, 1978), which is often used in other studies. As a large amount of the results concerns the error rates for 6 sample sizes n_i = 5,...,50, it seems reasonable to allow a couple of exceedances within this range.

Tables

The results of the simulation study described above are reproduced in a number of tables. They report the proportions of rejections of the corresponding null hypothesis, for different models and n_i = 5,10,..,50, small and large designs as well as balanced and unbalanced designs, where for the tests of the type I error α=0.05 has been chosen. They are available online (see address below):

- D 1: type I error rates of the ANOVA tests for fixed n_i ,
- D 2: type I error rates and power of the tests for homogeneity of covariances in relation to n_i.
- D 3: type I error rates and power of the tests for homogeneity of correlations in relation to n_i.

All references to these tables will be referred as D *n.n.n*. They can be viewed online:
http://www.uni-koeln.de/~luepsen/statistik/texte/comparison-tables/.
The most important of them are summarized in the appendix.

4. 1 ANOVA tests: Type I error

The univariate F test keeps the type I error rate of the main effect of factor B completely under control, for all heterogeneity models, distributions and situations considered, while the multivariate tests show slight exceedances in unbalanced large designs for n_i=5 with type I error rates between 0.75 and 0.8 for a nominal α=0.05 (see appendix table 2 and D 1). As Algina (1994) and Keselman et al. (1993) pointed out, the test of the interaction is more critical. But there is another displeasing situation in which all methods tend to show excessive type I error rates for the test for the main effect of factor A. The simulation results for these two effects will be described below.

The univariate F test

The type I error rates of the F test for the interaction effect are mainly affected for unbalanced designs which is also the result from the empirical literature (e.g. Huynh, 1978, and Keselman et al., 1993), though there are some slight exceedances also for balanced designs, predominantly for smaller sample sizes, with rates up to 0.09 in case of a univariate distribution, actually if there is no pairing (see table 2 and D 1.1.2). Therefore only the results for the interaction effect in unbalanced designs will be discussed, first model 1. Even if there is no pairing, the error rates rise up to 0.085 for the F test (for all distributions considered), particularly for lower n_i in small designs, which is hardly reduced by the Huynh-Feldt adjustment (see table 2). In case of a mo-

derate negative pairing the error rates exceed the interval of liberal robustness for all distributions, and rise up to 0.093 even for an underlying multivariate normal distribution. In situations of severe negative pairing the type I error is completely out of control with values beyond 0.20 for all distributions. Things are vice versa in the case of positive pairing leading to a very conservative behavior which is well-known from between subject designs.

Next model 2: as long as there is no relation between the correlations $r^{(i)}$ within group i and the sample size n_i the F test keeps the nominal error level (see table 2 and D 1.5.2). In case of a large positive relation between $r^{(i)}$ and n_i the F test offends the interval of liberal robustness for all distributions with rates up to 0.16. If there is a large negative relation between $r^{(i)}$ and n_i its reaction is reverse and leads to a conservative behavior. Concerning model 3 there are no remarkable exceedances for the F test.

Finally model 4, where the conditions of models 1 and 2 are combined. From the previous results it had to be expected that a negative pairing of the variances together with a positive pairing of the group correlations leads to exploding type I error rates of the univariate F test, with values rising to 0.40 and beyond (see table 2 and D 1.10.2), and that a positive pairing of the variances together with a negative pairing of the group correlations leads to extreme conservative results, with rates below 0.01 for a nominal α=0.05 (see table 2 and D 1.11.2). In the other two combinations the effects cancel each other out. It should be noted that generally the Huynh-Feldt adjustment is hardly able to reduce exceeding error rates. Concerning the differences of the type I error rates between small and large designs, extreme reactions are all in all more damped in small designs.

Probably a surprising result: a negative relation between the correlations $r^{(i)}$ within group i and the sample size n_i leads to excessive type I error rates for the test of the grouping factor in unbalanced designs. In case of a large pairing (corr=-0.9) the rates lie between 0.07 and 0.09 for small designs and between 0.107 and 0.12 for large designs for a nominal α level of 0.05, independent of the sample sizes (see table 2 and D.1.7).

Concerning the impact of the structure of the within group correlation matrices there are a couple of studies reporting an increase of the type I error rates (see e.g. Harwell & Serlin, 1994). The same can be observed in this study for the situations in model 1, where the type I error rates lie about 10-20 percent above those for equal correlations. Its behavior is converse in the situations of model 2. Here unequal correlations have a dampening effect on the exceeding error rates, while the maximum is reached for equal correlations within a group, but differing between the groups. This explains also the behavior of the F test in model 3 with unequal and varying within group correlations $r^{(i)}$ for i=1,..,I, where the F test did not show any conspicuousness.

The multivariate test

The results for the multivariate tests are to a large extent dependent from the choice of test: Hotelling-Lawley‘s test behaves generally rather liberal, whereas Pillai‘s test reacts more conservatively. And the error rates of Wilks‘ test lie nearly exactly in between. Concerning the interaction effect in designs with equal n_i the tests by Pillai and Wilks keep the type I error completely under control, as known from the empirical literature quoted above. In contrast the one by Hotelling-Lawley, which generally exhibits increased type I error rates in large designs for $n_i \leq 10$ with values between 0.07 and 0.10, for all distributions and for all the heterogeneity models. Next the interaction effect in unbalanced designs will be discussed. The appearance is here similar to the one described for equal sample sizes: Hotelling-Lawley‘s test shows exceeding type I error rates in large designs for $n_i \leq 20$, while Pillai‘s test performs almost exactly

like the F test. Particularly in cases where the F test cannot control the type I error, neither the multivariate tests can do it.

With regard to the effect of a correlation structure with unequal within group correlations the multivariate tests behave similarly as the F test with type I error rates which lie about 10-20 percent above those for equal correlations. Finally it should be reminded that there is no specific multivariate test for the effect of the grouping factor.

The nonparametric test

Although the tables of appendix D1 comprise the results for all three nonparametric methods, here only the behavior of the nonparametric generalized Kruskal-Wallis and Friedman test (KWF) is discussed, as this one is the most benevolent. It reveals a perfect type I error control in nearly all situations (see table 2). The error rates are kept even in the moderate interval of robustness for all underlying distributions as well as for all four models, except for an underlying mixed skewed distribution. In such cases predominantly the test of the repeated measures main effect is affected: for model 1 and 3 with error rates up to 0.09 for a nominal α level of 0.05 (see e.g. D 1.1.1 and D 1.3.1) and for model 2 up to 0.16 (see e.g. D 1.7.1). In case of models 2 and 4 also the test of the interaction is concerned. For model 2 and a positive pairing of n_i and $r^{(i)}$, as well as for all situations of model 4 the type I error rates rise up to 0.17 for increasing n_i ->50 (see e.g. table 2).

However, for model 2 and a negative pairing of n_i and $r^{(i)}$ the type I error for the test of the grouping main effect is no more completely under control, similarly to the behavior of the F test. Here the results depend not only on the design, but to a great extent on the underlying distribution. In case of a normal distribution the rates are kept in the interval of liberal robustness. The same applies to the lognormal, uniform and mixed skewed distribution except for large unbalanced designs, where the rates rise up to 0.09. For these distributions the type I error control is much better than for the F and multivariate tests. But in case of the exponential distribution the error rates increase up to 0.09 for small designs and up to 0.18 for large designs, which lies beyond the values of the parametric tests (see table 2 and D 1.7.1).

Concerning the impact of the correlation structure the generalized Kruskal-Wallis and Friedman test presents itself robust to unequal correlations, in contrary to the parametric tests.

4. 2 Tests for homogeneity of covariance matrices

Type I error

The LR test has generally inflated type error rates for small n_i: in case of underlying normal distributions only for $n_i \geq 50$ they stay in an acceptable range, for uniform und mixed skewed distributions they need about $n_i \geq 20$. For strongly skewed distributions the type I error behavior is completely unacceptable, with rates generally above 0.20 for the lognormal and above 0.70 for the exponential distribution at a nominal level of α=0.05. Somewhat better the type I error control of Box‘s test: in case of the normal, uniform and mixed skewed distributions the rates stay in the interval of liberal robustness for $n_i \geq 10$, except the situation of large unbalanced designs, where $n_i \geq 30$ is needed, whereas for strongly skewed distributions the Box test, too, has unacceptable error rates above 0.10. The type I error control is the same for equal as well as for unequal correlations. Generally the empirical error rates of Box‘s test lie considerably below those of the LR test, particularly for small n_i. For details see table 3 and D 2.1.

Schott‘s tests T_1, T_2 and T_3 score rather variably, though the type I error rates are fairly similar

for T_2 and T_3. T_1 keeps the error rates completely under control for normal, uniform and mixed skewed distributions, similar to Box‘s test, whereas the test shows exorbitant rising rates for increasing n_i in case of skewed distributions, up to 0.15 (n_i=50) for the lognormal distribution and up to 0.60 for the exponential distribution. Totally different Schott‘s T_3: the test cannot control the type I error in the cases of the uniform and the mixed skewed distributions with rates up to 0.16 (for T_2) and 0.60 (for T_3), but reveals acceptable rates in case of the strongly right skewed distributions, at least for n_i≥10 (balanced design) and n_i≥20 (unbalanced design). Comparing the results for equal and unequal correlations, the behavior is rather similar here, too.

The only tests which control the type I error in nearly all situations are the Levene test and the robust dispersion test RD. The error rates of the first one exceed only sometimes the limit of liberal robustness for n_i=5 in unbalanced designs, while those of the second one rise beyond the limit for n_i≥30 in case of an underlying discrete uniform distribution (see appendix table 3 and D2).

The power

First it has to be stated, that, not surprisingly, a considerable number of excellent power performances of the LR, Box and Schott‘s tests occur in those situations where they are not able to control the type I error. Therefore only those settings are regarded, in which the tests considered show acceptable error rates. This study confirmed the well-known experience that the performance of the tests, not only of the Likelihood ratio and Box M test, but also each of the three tests by Schott, is extremely dependent on the underlying distribution. As had to be expected, the LR and Box M test show overall the best power rates from all seven methods in case of an underlying normal distribution, and for every of the three models for heterogeneous covariance matrices. However, due to the poor type I error control the LR test requires sufficiently large n_i≥50. Generally the LR test scores slightly better than the M test. And on the other side, the Levene test, which has the perfect type I error control, cannot always keep up with the other tests. Let us have a detailed look how the different tests score in the three models.

First model 1, where the groups differ in respect of the variances, and the case of no pairing. The best performing method in all situations is the RD procedure. But the other methods, too, show excellent power rates, as long as they keep the type I error under control. For an underlying normal distribution and small $n_i \leq 10$ Schott‘s T_3 test has the best power while for larger n_i these are the Levene, T_1, T_2 and Box test. In case of a lognormal distribution the Levene and Schott‘s T_2 and T_3 tests achieve very similar rates, rising up to 1.00 (for n_i=50). For uniform, exponential and mixed skewed distributions Levene‘s test is definitely the most attractive, because either Schott‘s T_2 and T_3 or Schott‘s T_1 are not applicable, and the rates of the remaining tests stay below those of Levene‘s test, and the LR and Box test are only valid for n_i≥30 or n_i≥20 respectively (see table 4 and D 2.3). Things look clearly different in unbalanced designs with either negative or positive pairing. In both situations the Levene and the RD procedures loose a considerable part of their power, while the other methods are able to improve their power. For example, if the correlation between the coefficients c_i (equivalent to the variances $s_{jj}^{(i)}$) and the sample sizes n_i is 0.7, then for n_i=20 the power of the Levene test decreases from 0.35 to 0.07 and for the RD from 0.97 to 0.57. Similar rates for a negative correlation. For details see table 4 as well as D 2.4 and D 2.5.

Next to model 2, where the correlations of the variables are different for the I groups but equal within each group, and first the case where correlations $r^{(i)}$ and sample sizes n_i are independent. A couple of noticeable and surprising peculiarities: in general the rates are about 50 percent lower than in model 1, the rates for large designs lie approximately 20 percent below those from

small designs, and the condition of unequal cell counts in large designs shows the smallest power rates. First the case of an underlying normal distribution. For small designs the Levene test is in front of Schott‘s tests, whereas it is vice versa for large designs. The Box test shows power rates above average, in particular for large designs, as well as LR test for large $n_i \geq 50$. Also for lognormal distributions Levene‘s test lies in front of Schott‘s T_3 and T_2 tests, but only for small designs, whereas for large designs the T_3 method performs the best. Similar results for the uniform distribution: Levene‘s test scores for small designs, in front of the LR, Box and T_1 methods, while for large designs the LR and Box test are leading. In case of the exponential distribution the Levene test again is leading clearly in all situations, whereas for the mixed skewed distribution this is the case only for small designs. For large ones the LR test, followed by Box‘s test, are definitely better, while Schott‘s T_1 has the lowest power (see table 4 and D 2.6). Things are considerably different in unbalanced designs if there is a relation between the correlations $r^{(i)}$ within the groups and sample sizes n_i. No matter if the correlation is positive or negative, the power of all tests is much higher than in the case of independence, except the Levene test, which reveals in this situation power rates mostly below 0.10 (see table 4 and D 2.7). The relation between the other methods is the same as described above.

Finally model 3, where the correlations of the variables differ not only from group to group but also within each group. The distribution of the $r_{j_1 j_2}^{(i)}$ has been chosen in a way that their group averages differ not as much as the $r^{(i)}$ do in model 2, in order to produce a different setting. In contrary to model 2, the rates for small designs lie here about 50 percent below those of large designs. And one more surprising peculiarity compared with the results for models 1 and 2: the power rates of Levene‘s test for small designs stay always below 0.06, i.e. this test has no power at all for small designs (see table 4 and D 2.9). For normal distributions the LR and Box tests are the best performing for $n_i \geq 50$. For smaller n_i these are Schott‘s tests, with T_3 in front of T_2 and T_1, while Levene‘s test is disappointing. The same applies to the situation of lognormal distributions. Similar results for the case of an uniform distribution. For exponential distributions the T_2 test is in front of Levene‘s test for small designs, whereas it is vice versa for large designs. Finally the situation of a mixture of skewed distributions: for small designs the LR, the Box and the T_1 methods are almost equal up, while for large designs the LR and the Box tests have a far better power than Schott‘s T_1. Levene‘s test cannot keep up in this situation, even not for large designs.

The results from model 2 and 3 reveal that none of the procedures for testing the homogeneity of covariance matrices is able to detect heterogeneity on a reliable level, if only the correlations are heterogeneous while the variances are equal, perhaps with one exception: when there is a strong relation between the correlations $r^{(i)}$ within the groups and the sample sizes n_i (see appendix table 4 and D2). Therefore additionally methods for comparing correlations matrices have to applied.

4. 3 Tests for homogeneity of correlation matrices

Type I error

Regarding the procedure by Jennrich the simulation results confirm the judgement of Larntz & Perlman (1985): this test needs large samples to render reliable results. For small $n_i \leq 15$ the error rates lie always far beyond the limit of robustness, often between 0.15 and 0.40 for a nominal α level of 0.05, they decrease with increasing sample sizes, but in many situations, particularly for large designs, they stay above the limit for n_i=50. Nearly the same results have been found for Box‘s test. In contrast, the Levene-like and the LP test behave far more benevolently: the first one controls the type I error for all distributions in all situations, while the second one

shows exceedances only for right skewed distributions, with rates up to 0.10 for $n_i \leq 20$ in case of large designs with underlying lognormal distribution, but unacceptable rates for the exponential distribution (see appendix table 5 and D3).

The power

As mentioned above both the Jennrich and the Box test have an insufficient type I error control in most situations, and on the other side there are only very few cases where their power rates exceed those of the Levene-like and the LP test. Therefore only the last two are discussed here. Generally the test by Larntz & Perlman scores better than the Levene-like test, often with power rates between 30 and 100 percent above. If there is a relation between the correlations within the groups $r^{(i)}$ and the sample sizes n_i the absolute power is comparatively high, often reaching 1.0 for n_i=50, and normally larger for large designs, but surprisingly rather low (about 25 percent below average) for small unbalanced designs. However, in the case of no correlation between $r^{(i)}$ and n_i the power rates reach only values between 0.5 and 0.8, but show rates about 20 percent above average just for small unbalanced designs (see appendix table 6 and D3).

5. Conclusions

This study confirms that the choice of a suitable method to test the equality of covariance matrices is largely dependent from the underlying distribution, but to the same extent from the pattern of heteroscedasticity, and also from the design, balanced or unbalanced, and small or large. This makes a general recommendation impossible. Nevertheless there are some regularities:

- Concerning the control of the type I error Levene‘s method is by far the best, with a complete control in all situations, followed by the RD procedure with violations only in one situation.
- From the design of the robust dispersion test, analyzing only the variances, it is evident that this method is not able to detect differences between the groups that arise from inequality of the correlations, as it is the case in models 2 and 3.
- The Levene and the RD tests are definitely the best choice as long as the correlations are equal for all groups and as long as there is no negative pairing between the group variances and sample sizes, because both have the largest power for all different underlying distributions.
- If the correlations $r^{(i)}$ are unequal for the different groups, the power level of the methods which compare the covariance matrices is generally low, and even worse for large designs where the rates stay below 0.25, and often even below 0.15.
- If the correlations $r^{(i)}$ are unequal for the different groups and additionally unequal within groups, the power level is generally low, too, but here the worse situation occurs for small designs, where the rates stay below 0.25, and often even below 0.15. Levene‘s test has the smallest power from all methods, for both, small and large designs.
- The likelihood ratio test is generally applicable only for larger samples, and even in the classical case of underlying multivariate normal distributions only for $n_i \geq 50$.
- Because of deficiencies of the type I error control there is none from Schott‘s methods which could be recommended for any distribution. The T_3 test is the least attractive, while the T_1 and T_2 are less suspicious. The T_1 test is to prefer for normal, uniform and mixed skewed distributions, whereas the T_2 test has advantages for right skewed distributions.
- The type I error control of Schott‘s T_3 test is generally better for balanced samples than for unbalanced. Therefore this method could be applied even in the case of skewed distribution if the cell counts are equal.

- Concerning a test of unequal correlation matrices the one by Larntz & Perlman is the only recommendable method, which has a satisfactory control of the type I error, though in case of large designs sometimes only for $n_i \geq 20$, and a very good power performance, except for model 3.

Beside those simulations described in section 3 and summarized in the tables of the appendix additional simulations has been performed:

- Nagao‘s approximation for the χ^2 test of Schott‘s T_1 yields no remarkable improvement.
- Using the nonparametric multivariate ANOVA by Agresti & Pendergast instead of the parametric Wilks Lambda for the computation of the Levene test yields similar rates for the type I error but considerably lower rates for the power.
- In case of an underlying uniform or exponential distribution the type I error and power rates have been observed for two conditions: continuous and discrete values of the outcome. Whereas this difference may have a substantial effect on the analysis of variance (see e.g. Luepsen, 2016), it seems to have nearly no remarkable impact on the comparisons of covariance matrices.

Table 1 gives an overview of the restrictions and power performances of the 7 methods for detecting unequal covariance matrices. But it has to be pointed to the fact that their behavior is more complex than that it could be reproduced in such a simple table. E.g. the relation between two methods changes sometimes for varying n_i, or there are differences between balanced and unbalanced designs. And the good results for the Levene test are restricted to the case of independence of group variances and sample sizes (no pairing). Therefore the table can only be a rough guide.

At the beginning it has been pointed to the fact that, as a check of an assumption for the analysis of variance, it is more reasonable to test the hypothesis $C\Sigma^{(1)}C‘ =...= C\Sigma^{(I)}C‘$ instead of the hypothesis $\Sigma^{(1)} =...= \Sigma^{(I)}$, where C is an orthonormal contrast matrix. But here, too, the study revealed that there is no substantial difference between the two approaches. However, the orthonormal transformation is occasionally able to reduce the problems mentioned above at least a bit. Thus the type I error rates are reduced, e.g. for Box‘s test and underlying normal or mixed skewed distributions, or for the T_2 test and underlying lognormal or exponential distributions, so that the rates stay in the interval of robustness generally for $n_i \geq 10$ (instead of $n_i \geq 20$). In regard of the power the differences between the two methods are not remarkable, sometimes positive and sometimes negative, and much depending on the methods.

Now to the final question: which method to choose? There is much criticism in the literature in regard to the Box M test and its sensibility for deviations from the multivariate normal distribution (e.g. Layard, 1974). But all in all this test is not worse than the other tests, too. For underlying normal, uniform and mixed skewed distributions its performance is above average, whereas for strongly skewed distributions its application is not recommendable. But the other methods show similar deficiencies. However, Schott‘s T_2 may be a reasonable supplement to Box‘s test for the situation of skewed distributions. And the Levene test leaves a good impression at first sight, but its power is to a large extent dependent from the equality of the correlation matrices and the pairing of group variances and sample sizes. Thus, if the distribution of the outcome *y* is known, either Box‘s test or Schott‘s T_2 are a good choice dependent on the underlying distribution. But if there is only little knowledge about the distribution of *y* - and this will be the normal case - one has to choose a method which is not sensitive to departures from the normal distribution. And looking at the results for the type I error control (see e.g. table 3), the Levene and the RD procedures are the tests of choice. Both show a perfect performance in case of he-

terogeneous variances, but poor power rates in case of unequal correlations, as the other methods, too. Therefore a nonsignificant p value may rise the question: is this due to homogeneity of the covariance matrices or due to unequal correlation coefficients $r^{(i)}$ for the groups? Hence, if the hypothesis of equal covariance matrices is not rejected, it is advisable to apply also a test for equal correlation matrices, where the test by Larntz & Perlman seems to be the most reliable.

model	distribution		restrictions to n_i (due to insufficient type I error control)							ranking
			LR	Box	T_1	T_2	T_3	Lev.	RD	
1	normal	eq ne	≥50 ≥60	≥10 ≥20						RD >LR >Box >Lev > (T_1,T_2,T_3)
	lognormal	eq ne	--	--	--	≥10 ≥30				RD > (Lev,T_2,T_3)
	uniform	eq ne	≥15 ≥20	≥10 ≥10		--	--		≤ 20	RD > (Lev, LR) > Box > T_1
	exponential	eq ne	--	--	--	≥20 ≥20	--	≥10		(RD, Lev) > T_2
	mixed skewed	eq ne	≥10 ≥30	≥10 ≥20		--	--	≥10		RD > Lev > LR > Box > T_1
2	normal	eq ne	≥50 ≥60	≥10 ≥20						small: (LR,Box,Lev) > (T_1,T_2,T_3) large: (LR,Box) > T_3 > (T_1,T_2) > Lev
	lognormal	eq ne	--	--	--	≥10 ≥30				small: Lev > T_3 > T_2 large: T_3 > (T_2,Lev)
	uniform	eq ne	≥15 ≥20	≥10 ≥10		--	--		≤ 20	small: Lev > (LR,Box) > T_1 large: LR > Box > Lev > T_1
	exponential	eq ne	--	--	--	≥20 ≥20	--	≥10		Lev >> T_2
	left/right skewed	eq ne	≥10 ≥30	≥10 ≥20		--	--	≥10		Lev > LR > Box > T_1 LR > Box > (Lev, T_1)
3	normal	eq ne	≥50 ≥60	≥10 ≥20						LR > Box > T_3 > T_2 > T_1>> Lev
	lognormal	eq ne	--	--	--	≥10 ≥30				T_3 > T_2 >> Lev
	uniform	eq ne	≥15 ≥20	≥10 ≥10		--	--		≤ 20	small: LR > (Box, T_1) >> Lev large: LR > Box > T_1 > Lev
	exponential	eq ne	--	--	--	≥20 ≥20	--	≥10		small: T_2 > Lev large: Lev >> T_2
	left/right skewed	eq ne	≥10 ≥30	≥10 ≥20		--	--	≥10		small: (LR, Box, T_1) > Lev large: LR > Box > T_1 > Lev

Table 1: restrictions (based on the type I error control) and rankings (concerning the power) for the 7 methods, 3 models, balanced and unbalanced designs.
"--" *means „not applicable", (.,.) means a "similar behavior", ">" means "better than" and ">>" means much "better than".*
For model 1 the rating of the Levene test is only valid if there is no distinct pairing, and for model 2 only if there is no distinct relation between group correlation $r^{(i)}$ *and cell counts* n_i.

Finally some practical aspects in conjunction with an ANOVA will be considered. If the parametric F test is planned, a test on homogeneity of the covariance matrices is only necessary for unbalanced designs. For these conditions section 4.1 revealed two critical situations: (a) negative pairing (negative correlation between group variances $s_{jj}^{(i)}$ and sample sizes n_i), (b) a positive correlation between $r^{(i)}$ and n_i and (c) a negative correlation between $r^{(i)}$ and n_i. Case (a) is a bit tedious: for each variable j=1,...,J the group variances $s_{jj}^{(i)}$ have to be computed and related to the sample sizes n_i. But since even small correlations may have an impact on the F test of the interaction effect, it seems reasonable to perform a test on equal covariances in any case. Case (b) is comparatively simple: for each group an average of the correlations of the upper (or lower) off-diagonal correlation matrix has to be computed as an estimate of $r^{(i)}$, and these have to be related to the sample sizes n_i. If this correlation is approximately zero or negative, the F tests for the interaction effect are not affected, i.e. there is no need to check the homogeneity of the correlations matrices. Case (c) works similar to case (b). If this correlation is approximately zero or positive, the F tests for the grouping main effect are not affected.

If a multivariate test is planned, also balanced designs have to be taken into account. Section 4.1 revealed, here too, two critical situations: (a) no pairing or positive pairing (positive correlation between group variances $s_{jj}^{(i)}$ and sample sizes n_i) and (b) a negative correlation between $r^{(i)}$ and n_i. Concerning case (a) it seems reasonable to perform a test on equal covariances in any case with the same argument as noted above. Case (b) can be treated similarly as described above for the F test.

At last it should be reminded that the nonparametric generalized Kruskal-Wallis and Friedman test (KWF) might be a viable and easy alternative. As mentioned in section 4.1 it has a perfect type I error control for all kinds of heterogeneous covariance matrices. Additionally it has the advantage not to require sphericity of the covariance matrices, though a test on equality of the variances is recommendable, e.g. a Levene-like test as proposed by Wilcox (1989). The only disadvantage: the power does not reach always that of the parametric F test, particularly for small sample sizes $n_i \leq 30$ and in the cases of the normal or exponential distributions with a loss of 10 to 20 percent, while for various kinds of nonnormal distributions and the situations of heterogeneous variances the KWF procedure has even a larger power (see Luepsen, 2020).

6. Programming

The simulations for this study have been programmed in R (version 3.5.3). For the data generation the function `mvrnorm` from the package `MASS` (see Ripley, 1987) has been used to receive multivariate normal distributed variates. Other multivariate distributions were obtained by suitable data transformations. Various functions have been chosen to analyze the simulated data: the function `aov` in combination with `drop1` (to receive type III sum of squares estimates in the case of unequal cell counts) for the standard ANOVA F-test, the function `lm` together with `anova` for the multivariate test by Hotelling & Lawley, an own function `np.anova` for the nonparametric ANOVAs (generalized Kruskal Wallis and Friedman test, van der Waerden test and Koch‘s procedure). For the tests on homogeneity of covariance matrices and correlation matrices also own functions `check.covar` and `check.corr` have been applied. For the own functions see Luepsen (2014).

Some of the computations have been performed on a Windows notebook, but for the major part the high performance cluster CHEOPS of the Regional Computing Centre (RRZK) of the university of Cologne has been used. I would like to thank the staff of the RRZK for their technical support as well as Prof. Unkelbach for his organizational support.

7. References

Agresti, A. & Pendergast, J. (1986): Comparing mean ranks for repeated measures data. *Communications in Statistics - Theory and Methods,* 15, No 5, pp 1417-1433.

Algina, J. (1994): Some Alternative Approximate Tests for a Split Plot Design. Multivariate Behavioral Research, 29 (4), pp.365-384).

Beasley, T. Mark; Sheehan, Janet K. (1994): *Choosing a MANOVA Test Statistic When Covariances Are Unequal*. Paper presented at the Annual Meeting of the Midwestern Educational Research Association, Chicago, IL.

Boik, Robert .J (1981): A priori tests in repeated measures designs: Effects of nonsphericity. *Psychometrika*, Vol 46, No 3, pp 241-255.

Box, G.E.P. (1949): A General Distribution Theory for a Class of Likelihood Criteria, Biometrika, Vol. 36, No. 3/4, pp. 317-346

Box, G.E.P. (1954): *Some theorems on quadrative forms applied in the study of analysis of variance problems, I: Effect of inequality of variance in the one-way classification.* Annals of Mathematical Statistics, 25, pp 290-302

Bradley, J.V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology*, 31, pp 144-152.

Emrich L.J. , Piedmonte M.R. (1992): On some small sample properties of generalized estimating equation estimates for multivariate dichotomous outcomes. *Journal of Statistical Computation and Simulation*, 41, 19-29 .

Feir, B.J., Toothaker, L.E. (1974). *The ANOVA F-Test Versus the Kruskal-WallisTest: A Robustness Study*. Paper presented at the 59th Annual Meeting of the American Educational Research Association in Chicago, IL.

Fouladi, R.T. & Yockey, R.D. (2002): Type I error control of two-group multivariate tests on means underconditions of heterogeneous correlation structure and varied multivariate distributions. *Communications in Statistics - Simulation and Computation*, Volume 31, Issue 3, 375-400

Greenstreet, R.L. & Connor, R.J. (1974): Power of Tests for Equality of Covariance Matrices, *Technometrics*, Vol. 16, No. 1, pp. 27-30

Hallin, Marc & Paindaveine, Davy (2009): *Optimal tests for homogeneity of covariance, scale, and shape.* Journal of Multivariate Analysis ,100, pp 422-444.

Harwell, M.R. & Serlin, R.C. (1994): A Monte Carlo study of the Friedman test and some competitors in the single factor, repeated measures design with unequal covariances. *Computational Statistics & Data Analysis,* 17, pp 35-49.

Huynh, H. (1978): *Some approximate tests for repeated measurement designs*, Psychometrika 43, 161-175

Jennrich, Robert I. (1970): An Asymptotic chi2 Test for the Equality of Two Correlation Matrices. Journal of the American Statistical Association, Vol. 65, No. 330, pp 904-912

Keselman, H. J., Carriere, K. C., & Lix, L. M. (1993): *Testing Repeated Measures Hypotheses*

When Covariance Matrices are Heterogeneous. Journal of Educational and Behavioral Statistics, Vol. 18, no. 4, pp 305-319

Keselman, H. J., Algina, J., Wilcox, R.R., Kowalchuk, R.K. (2000): Testing Repeated Measures Hypotheses when Covariance Matrices are Heterogeneous: Revisiting the Robustness of the Welch-James Test again. *Educational and Psychological Meaasurement*, Vol 60, No. 6, pp 925-938

Koch, Gary (1969): Some aspects of the statistical analysis of split plot experiments in completely randomized layouts. *Journal of the American Statistical Association*, Vol. 64, No. 326, pp. 485-505

Larntz, Kinley & Perlman, Michel D. (1985): A simple Test for the Equality of Correlation Matrices. University of Washington, Technical Report No. 63.

Layard, M.W.J. (1974): A Monte Carlo comparison of tests for equality of covariance matrices, *Biometrika*, 16, 3, pp 461-465

Luepsen, H. (2014): *R Functions for the Analysis of Variance*. URL: http://www.uni-koeln.de/~luepsen/R/ .

Luepsen, H. (2016): The aligned rank transform and discrete variables: A warning, *Communications in Statistics - Simulation and Computation*, DOI: 10.1080/03610918.2016.1217014

Luepsen, H. (2020): Some rank based ANOVA procedures for analyzing data from split-plot designs. To be published.

Marden, John, Gao, Yonghong (2002): Rank-based Procedures for Structural Hypotheses on Covariance Matrices. The Indian Journal of Statistics, Vol. 64, Series A, Pt 3, pp 653-677.

Nagao, Hisao (1973): On Some Test Criteria for Covariance Matrix. *The Annals of Statistics*, Vol. 1, No. 4 (Jul., 1973), pp. 700-709

O'Brien, Peter C. (1992): Robust Procedures for Testing Equality of Covariance Matrices. *Biometrics*, Vol. 48, No. 3 (Sep., 1992), pp. 819-827

Pervaiz, M.K. & Skinner, C.J. (1990): A Monte Carlo Comparison of Elliptical-Theory and other Tests for Equality of Covariance Matrices. Australian and New Zealand Journal of Statistics, Vol 32, 1, pp. 71-86

Peterson, K. (2002). Six Modifications Of The Aligned Rank Transform Test For Interaction. *Journal Of Modem Applied Statistical Methods*. Vol. 1, No. 1, pp 100-109.

Ripley, B. D. (1987): Stochastic Simulation. Wiley, New York.

Schott, J.R. (2001): Some tests for the equality of covariance matrices. *Journal of Statistical Planning and Interference*, 94, pp 25-36.

Wilcox, Rand R. (1989): Comparing the Variances of dependent Groups. Psychometrika, vol 54, No. 2, pp 305-315

Zaiontz, Charles: *Box's M Test Basic Concepts*, in Real Statistics Using Excel, URL: https://www.real-statistics.com/multivariate-statistics/boxs-test/boxs-test-basic-concepts/

Appendix

Table 2: Type I error rates of the ANOVA procedures

	cell counts	equal								unequal							
	design	small				large				small				large			
model	distribution	F	HL		KWF	F	HL		KWF	F	HL		KWF	F	HL		KWF
$\Sigma^{(i)} = c_i\Sigma^{(1)}$ $r(c_i,n_i)\sim 0$	normal	6.30	8.05	6.05	5.30	7.00	10.80	6.75	5.35	6.35	6.35	6.05	5.20	7.30	12.05	6.15	5.05
	lognormal	6.20	6.45	6.20	5.10	7.30	9.90	6.80	5.45	6.35	6.60	6.00	5.45	6.65	11.50	6.45	4.70
	uniform	6.55	7.45	6.30	5.20	6.95	10.20	6.45	5.10	6.35	6.60	6.15	5.25	7.90	12.50	6.45	5.80
	exponential	6.00	6.15	5.80	5.30	6.80	7.45	6.20	5.35	6.65	6.20	5.95	5.20	7.30	9.70	5.85	5.05
	mix skewed	6.60	7.65	5.90	6.15	6.80	9.45	6.60	6.85	6.60	6.30	5.90	6.20	7.10	11.45	6.15	6.70
$\Sigma^{(i)} = c_i\Sigma^{(1)}$ $r(c_i,n_i)\sim -0.3$	normal	6.40	7.80	5.85	5.30	6.70	10.25	6.45	5.35	7.95	7.85	7.80	5.20	9.00	13.10	8.80	5.05
	lognormal	6.05	6.55	6.20	5.10	6.90	9.00	6.30	5.45	8.15	8.25	8.05	5.45	9.15	12.95	8.90	4.70
	uniform	6.40	6.55	5.70	5.50	6.40	9.05	6.15	5.50	7.80	8.35	7.40	5.80	9.25	13.35	9.05	5.95
	exponential	6.30	6.15	5.80	5.30	6.30	6.95	5.90	5.35	8.10	8.00	7.70	5.20	8.90	11.30	7.90	5.05
	mix skewed	6.65	7.35	5.70	6.15	6.05	8.75	6.00	6.85	8.10	8.10	7.65	6.20	8.75	12.80	8.65	6.70
$\Sigma^{(i)} = c_i\Sigma^{(1)}$ $r(c_i,n_i)\sim -0.7$	normal	6.35	6.65	5.45	5.30	7.60	9.90	6.45	5.35	11.65	12.25	11.00	5.20	22.30	22.75	21.25	5.05
	lognormal	6.20	6.55	6.35	5.10	7.10	9.75	6.55	5.45	12.05	12.05	11.40	5.45	22.25	23.15	20.95	4.70
	uniform	6.25	6.80	5.60	5.55	7.00	9.95	6.75	5.40	11.40	11.75	11.20	5.65	22.35	22.85	21.25	5.90
	exponential	6.05	6.30	6.15	5.30	6.80	7.30	6.45	5.35	11.20	11.60	11.40	5.20	21.75	21.00	18.55	5.05
	mix skewed	6.40	7.00	5.70	6.15	6.50	9.85	6.75	6.85	12.00	12.05	11.05	6.20	22.50	23.20	21.35	6.70
$\Sigma^{(i)} = c_i\Sigma^{(1)}$ $r(c_i,n_i)\sim 0.7$	normal	6.35	7.30	6.65	5.30	6.40	9.25	6.60	5.35	4.10	5.10	2.85	5.20	1.80	3.30	1.45	5.05
	lognormal	6.80	6.75	6.60	5.10	6.75	8.90	6.70	5.45	3.45	3.70	3.60	5.45	2.40	3.80	1.80	4.70
	uniform	6.10	7.15	6.15	5.20	6.90	9.45	6.30	5.25	3.75	5.15	2.80	4.60	1.90	3.50	1.70	4.90
	exponential	5.75	6.20	5.90	5.30	6.50	7.55	7.00	5.35	3.30	3.65	2.85	5.20	2.10	3.10	1.40	5.05
	mix skewed	6.50	7.05	6.35	6.15	6.55	9.30	6.15	6.85	4.05	4.90	3.05	6.20	1.60	3.30	1.80	6.70
unequal correlations $r^{(i)}$ $r(r_i,n_i)\sim 0$	normal	6.5	6.50	6.40	6.00	6.10	7.05	5.85	5.55	5.85	6.90	5.35	5.30	6.60	8.00	6.45	5.10
	lognormal	6.1	6.25	6.10	6.00	6.05	6.25	5.45	5.55	6.05	6.70	5.45	5.30	6.85	8.20	6.25	5.10
	uniform	6.4	6.35	6.05	5.45	5.65	7.15	5.45	5.90	5.60	6.45	5.45	4.95	6.60	8.10	6.05	5.30
	exponential	5.4	5.50	5.55	6.00	5.95	5.95	5.30	5.55	6.70	6.05	5.90	5.30	7.35	6.80	6.45	5.10
	mix skewed	6.5	6.30	6.10	8.20	6.05	6.90	5.55	7.60	6.25	6.60	6.00	7.70	6.85	8.65	7.00	6.10
unequal correlations $r^{(i)}$ $r(r_i,n_i)\sim 0.9$	normal	6.40	7.05	6.35	5.75	6.80	7.80	6.55	5.85	12.20	12.85	11.85	5.05	17.65	19.55	17.00	5.25
	lognormal	6.25	6.25	6.20	5.75	6.80	7.60	6.10	5.85	12.10	12.60	11.95	5.05	17.20	18.75	16.95	5.25
	uniform	6.05	6.75	5.95	5.35	6.15	7.90	6.05	5.35	11.00	11.15	10.90	6.25	16.00	17.85	15.55	7.40
	exponential	5.50	5.65	5.45	5.75	7.05	6.75	5.75	5.85	11.05	11.60	11.15	5.05	16.55	16.55	15.30	5.25
	mix skewed	6.20	6.90	6.15	8.75	7.00	8.75	6.50	18.25	12.05	12.05	11.75	8.55	17.05	18.65	16.55	17.55
unequal correlations $r^{(i)}$ $r(r_i,n_i)\sim -0.9$	normal	6.50	6.20	6.15	5.75	6.15	8.75	6.25	5.50	3.80	4.10	3.30	5.05	1.95	3.35	1.75	5.05
	lognormal	6.75	6.80	6.60	5.75	5.90	7.75	6.10	5.50	3.30	4.00	3.05	5.05	2.00	3.30	1.75	5.05
	uniform	5.65	6.15	5.50	5.50	6.45	7.85	6.10	5.80	3.70	4.50	3.25	4.40	2.30	3.00	2.20	4.30
	exponential	6.35	6.70	6.50	5.75	5.90	6.65	5.65	5.50	3.55	3.75	3.45	5.05	2.30	2.85	2.05	5.05
	mix skewed	6.65	6.60	6.30	5.45	6.45	8.70	6.10	17.75	3.90	4.45	3.25	5.35	2.00	3.45	1.85	17.00
varying correlations $r^{(i)}_{j1j2}$ for variables j_1, j_2 within each $\Sigma^{(i)}$	normal	6.45	6.35	5.95	5.55	6.15	7.10	5.70	5.35	6.60	7.50	7.25	5.40	6.40	6.70	5.95	5.50
	lognormal	6.15	6.25	6.00	5.55	6.25	6.65	5.85	5.35	6.40	7.00	6.60	5.40	6.50	6.60	5.80	5.50
	uniform	6.55	6.45	5.80	5.70	5.75	6.30	5.55	5.55	6.75	7.35	6.95	5.65	6.35	6.55	5.80	5.25
	exponential	5.50	5.80	5.55	5.55	6.35	6.20	5.90	5.35	6.50	6.55	6.05	5.40	6.40	6.35	5.30	5.50
	mix skewed	6.15	6.40	5.85	6.90	6.10	6.90	5.65	7.55	6.45	7.20	6.65	7.40	6.00	6.20	5.45	7.10
$\Sigma^{(i)} = c_i\Sigma^{(1)}$ $r(c_i,n_i) \sim -0.7$ combined with unequal corr $r^{(i)}$ $r(r_i, n_i)\sim 0.9$	normal	5.35	6.00	5.70	5.60	6.75	7.90	6.40	5.85	5.00	4.95	4.90	5.40	6.95	7.40	6.35	4.90
	lognormal	6.10	6.25	6.35	5.60	6.30	7.50	6.40	5.85	5.15	5.05	4.95	5.40	7.30	7.60	6.65	4.90
	uniform	5.80	5.70	5.35	5.10	6.25	7.95	6.50	5.40	5.05	5.35	4.80	5.20	7.50	8.25	6.85	4.65
	exponential	5.80	5.95	5.60	5.60	6.70	6.45	5.95	5.85	5.15	5.25	5.15	5.40	9.00	8.20	7.50	4.90
	mix skewed	6.00	6.00	5.70	5.45	8.65	10.00	8.95	18.85	5.45	5.45	5.15	5.05	10.10	10.65	9.50	16.80

	cell counts	equal								unequal							
	design	small				large				small				large			
model	distribution	F	HL		KWF	F	HL		KWF	F	HL		KWF	F	HL		KWF
$\Sigma^{(i)} = c_i\Sigma^{(1)}$	normal	6.95	7.95	7.15	5.60	9.55	16.45	8.25	5.20	22.75	22.50	22.05	5.40	42.15	45.95	40.40	5.00
$r(c_i,n_i)$~ -0.7	lognormal	7.00	8.60	7.30	5.60	8.70	15.50	8.55	5.20	22.75	22.50	21.40	5.40	41.25	44.15	40.30	5.00
combined with	uniform	7.65	8.95	7.15	5.65	8.80	16.55	8.50	5.60	21.15	21.55	20.85	5.70	40.25	45.25	39.50	6.85
unequal corr. $r^{(i)}$	exponential	7.25	7.70	7.40	5.60	9.10	12.80	8.70	5.02	20.15	20.55	19.95	5.40	39.95	40.85	38.55	5.00
$r(r_i, n_i)$~ 0.9	mix skewed	7.30	8.55	7.15	8.50	7.65	15.50	6.35	18.70	21.55	22.35	21.45	8.15	37.60	41.55	36.15	18.60
$\Sigma^{(i)} = c_i\Sigma^{(1)}$	normal	6.90	7.20	6.95	5.60	8.75	14.50	7.90	5.85	3.40	4.10	3.05	5.40	0.95	2.50	0.70	4.90
$r(c_i,n_i)$~ 0.7	lognormal	7.20	7.55	7.40	5.60	8.90	14.15	7.90	5.85	3.35	4.05	3.00	5.40	1.00	2.70	0.85	4.90
combined with	uniform	6.35	7.35	6.85	5.85	7.85	13.15	7.70	6.60	3.75	4.80	3.55	4.35	1.05	3.25	0.90	3.85
unequal corr. $r^{(i)}$	exponential	7.45	7.90	7.70	5.60	8.65	11.55	7.45	5.85	3.05	3.30	2.35	5.40	1.10	2.55	1.10	4.90
$r(r_i, n_i)$~ 0.9	mix skewed	6.75	7.20	6.85	5.45	10.40	17.75	10.10	18.85	3.15	4.50	3.05	5.05	1.35	3.75	1.30	16.80
$\Sigma^{(i)} = c_i\Sigma^{(1)}$	normal	6.95	6.55	6.35	5.60	7.10	9.75	6.55	5.20	4.70	5.50	4.10	5.40	7.70	10.85	5.55	5.00
$r(c_i,n_i)$ ~ 0.7	lognormal	6.45	6.45	6.35	5.60	6.60	8.60	6.50	5.20	4.50	5.45	4.05	5.40	8.35	10.75	5.45	5.00
combined with	uniform	6.20	6.20	5.85	5.50	6.55	9.80	6.55	5.00	4.65	5.25	4.25	5.10	6.85	10.70	5.50	5.65
unequal corr. $r^{(i)}$	exponential	5.70	5.90	5.70	5.60	7.15	6.85	6.35	5.20	3.90	3.95	3.85	5.40	9.30	9.00	4.80	5.00
$r(r_i, n_i)$~ 0.9	mix skewed	6.10	6.40	6.30	8.50	3.85	6.80	3.80	18.70	4.50	5.40	4.00	8.15	4.45	8.45	4.70	18.60

Table 2: Maximum of the type I error rates (in the range of n_i=5,..,50) for the univariate F test (F), the multivariate tests by Hotelling-Lawley (HL) and Pillai (Pillai) and the generalized Kruskal-Wallis-Friedman test (KWF) for 12 models of covariance heterogeneity, 5 distributions, small and large designs, equal und unequal cell counts,
above for the interaction effect, below for the main effect of the grouping factor

	cell counts	equal								unequal							
	design	small				large				small				large			
model	distribution	F	HL		KWF	F	HL		KWF	F	HL		KWF	F	HL		KWF
unequal	normal	6.20	6.20	6.20	5.9	6.05	6.05	6.05	5.55	5.85	5.85	5.85	6.0	5.15	5.15	5.15	5.25
correlations $r^{(i)}$	lognormal	5.85	5.85	5.85	5.8	5.95	5.95	5.95	6.25	5.95	5.95	5.95	5.8	5.15	5.15	5.15	5.75
$r(r_i,n_i)$~ 0	uniform	6.65	6.65	6.65	5.9	6.20	6.20	6.20	6.15	6.30	6.30	6.30	5.4	5.35	5.35	5.35	5.20
	exponential	5.85	5.85	5.85	6.3	5.85	5.85	5.85	8.20	5.65	5.65	5.65	6.5	5.45	5.45	5.45	8.20
	mix skewed	6.35	6.35	6.35	5.2	6.10	6.10	6.10	6.05	5.95	5.95	5.95	5.5	5.30	5.30	5.30	5.80
unequal	normal	5.85	5.85	5.85	5.85	6.35	6.35	6.35	5.95	5.15	5.15	5.15	4.50	3.50	3.50	3.50	3.70
correlations $r^{(i)}$	lognormal	5.60	5.60	5.60	5.80	6.20	6.20	6.20	7.00	5.10	5.10	5.10	4.80	3.60	3.60	3.60	4.85
$r(r_i,n_i)$~ 0.9	uniform	6.15	6.15	6.15	6.05	6.50	6.50	6.50	6.30	5.45	5.45	5.45	3.95	3.25	3.25	3.25	3.60
	exponential	5.95	5.95	5.95	8.80	6.10	6.10	6.10	15.55	5.30	5.30	5.30	5.60	3.75	3.75	3.75	11.00
	mix skewed	6.05	6.05	6.05	5.15	6.35	6.35	6.35	7.60	5.45	5.45	5.45	4.50	3.25	3.25	3.25	4.25
unequal	normal	6.35	6.35	6.35	5.50	7.05	7.05	7.05	6.30	8.85	8.85	8.85	6.50	11.00	11.00	11.00	8.05
correlations $r^{(i)}$	lognormal	5.95	5.95	5.95	6.25	6.90	6.90	6.90	7.55	8.60	8.60	8.60	6.60	11.15	11.15	11.15	9.10
$r(r_i,n_i)$~ -0.9	uniform	6.10	6.10	6.10	5.65	6.60	6.60	6.60	6.40	9.05	9.05	9.05	7.40	11.85	11.85	11.85	9.05
	exponential	5.75	5.75	5.75	8.85	6.95	6.95	6.95	16.35	8.25	8.25	8.25	8.85	11.55	11.55	11.55	18.10
	mix skewed	6.50	6.50	6.50	5.60	6.95	6.95	6.95	7.20	9.20	9.20	9.20	7.05	11.65	11.65	11.65	10.30

Table 3: Type I error rates of the tests for homogeneity of covariance matrices

	cell counts	equal				unequal			
	design	small		large		small		large	
distribution	method	10	50	10	50	10	50	10	50
normal	LR	11.70	5.65	47.90	6.65	14.90	7.50	82.00	9.60
	Box M	4.85	4.65	5.90	4.25	4.85	5.90	18.35	5.00
	Schott T1	2.25	3.85	1.60	3.60	2.50	5.50	3.30	4.60
	Schott T2	5.00	3.90	7.35	3.95	10.25	6.70	20.20	4.95
	Schott T3	9.50	3.75	14.85	4.10	22.35	5.70	40.75	5.45
	Levene	4.25	3.95	5.05	4.40	3.75	4.70	5.80	4.55
	R Dispersion	3.90	4.75	4.45	5.30	4.00	4.15	3.75	4.85
lognormal	LR	19.60	18.40	63.75	32.55	24.05	20.35	88.50	33.55
	Box M	9.30	16.60	12.65	26.05	10.00	17.15	25.45	24.30
	Schott T1	4.50	14.80	7.80	25.95	6.70	15.75	9.70	23.90
	Schott T2	5.50	5.20	5.25	4.95	9.80	6.85	15.05	5.10
	Schott T3	8.60	5.70	9.75	5.90	19.45	7.50	32.75	6.55
	Levene	4.45	4.05	5.65	4.80	4.85	4.85	5.20	4.60
	R Dispersion	3.30	4.35	3.00	3.20	2.85	3.50	3.95	3.30
uniform (discrete)	LR	5.20	1.05	27.30	0.85	7.10	2.00	62.30	1.80
	Box M	1.85	0.75	1.80	0.50	1.80	1.55	7.15	1.05
	Schott T1	0.60	0.70	0.25	0.30	0.55	1.10	0.70	0.70
	Schott T2	18.40	16.45	51.70	54.20	27.70	16.55	57.85	53.65
	Schott T3	33.55	42.00	66.80	67.75	47.55	40.75	76.50	67.15
	Levene	3.85	3.60	4.45	3.30	4.15	4.40	4.35	3.65
	R Dispersion	3.10	10.90	2.90	14.85	3.10	9.60	3.00	14.65
exponential	LR	65.20	71.90	97.50	96.40	66.95	71.50	99.30	95.65
	Box M	48.40	70.20	74.65	94.70	45.60	68.75	73.45	93.40
	Schott T1	36.10	67.70	64.70	94.30	31.65	63.60	53.10	88.75
	Schott T2	5.05	5.25	1.85	1.75	7.95	7.30	5.05	3.05
	Schott T3	8.45	9.75	6.20	4.55	17.30	12.00	18.95	6.50
	Levene	6.25	5.25	7.45	6.15	5.80	5.95	7.15	6.10
	R Dispersion	2.10	2.90	2.35	3.05	2.05	2.85	2.70	2.50
mixed skewed	LR	6.30	1.30	31.85	1.55	8.70	2.50	73.90	2.80
	Box M	1.80	0.95	2.65	0.80	2.95	2.00	13.50	1.40
	Schott T1	0.60	0.75	0.70	0.60	0.90	1.00	1.20	0.65
	Schott T2	13.25	12.05	40.20	41.00	22.35	13.70	49.45	40.55
	Schott T3	27.65	16.20	56.90	61.35	41.85	19.50	70.90	60.35
	Levene	4.50	3.45	4.85	4.35	4.90	3.65	4.95	4.30
	R Dispersion	2.15	3.85	3.35	3.75	2.55	3.45	2.25	3.25

Table 3: type I error rates for 7 methods of testing the homogeneity of covariance matrices, for n=10, 50, for 5 distributions, small and large designs, equal und unequal cell counts.

Table 4: Power of the tests for homogeneity of covariance matrices

	cell counts	equal				unequal			
	design	small		large		small		large	
model	method	10	50	10	50	10	50	10	50
$\Sigma^{(i)} = c_i\Sigma^{(1)}$ $r(c_i, n_i) \sim 0$	Box M	13.55	90.75	17.90	99.50	13.85	97.90	34.90	99.65
	Schott T1	7.35	88.15	14.20	99.50	4.55	96.95	17.00	99.45
	Schott T2	13.60	88.10	24.95	99.45	15.35	95.85	41.85	99.30
	Schott T3	19.30	82.75	34.10	93.90	27.30	93.20	52.65	94.10
	Levene	11.95	90.20	21.25	99.50	13.65	97.15	24.75	99.75
	R Dispersion	35.50	98.95	74.30	100.0	41.05	100.0	80.10	100.0
$\Sigma^{(i)} = c_i\Sigma^{(1)}$ $r(c_i, n_i) \sim -0.7$	Box M	23.15	99.60	39.85	100.0	19.70	96.80	46.40	99.95
	Schott T1	13.75	99.30	42.90	100.0	26.05	97.55	67.65	100.0
	Schott T2	21.15	99.10	48.35	100.0	37.85	97.05	75.10	100.0
	Schott T3	24.20	96.35	46.35	96.05	45.65	90.15	77.05	91.40
	Levene	4.05	29.05	4.20	59.00	5.00	56.50	6.00	50.25
	R Dispersion	6.05	54.05	19.20	99.65	10.00	81.75	16.90	93.80
$\Sigma^{(i)} = c_i\Sigma^{(1)}$ $r(c_i, n_i) \sim 0.7$	Box M	42.55	100.0	40.60	100.0	22.15	99.35	44.70	100.0
	Schott T1	26.80	100.0	37.10	100.0	7.40	93.35	3.75	99.95
	Schott T2	36.05	100.0	43.25	100.0	22.05	92.50	24.30	99.95
	Schott T3	33.15	98.55	45.55	96.25	32.55	79.15	45.00	91.15
	Levene	12.70	45.35	10.65	65.60	9.85	47.10	15.15	99.05
	R Dispersion	18.60	75.50	38.95	99.15	15.50	79.45	72.85	100.0
unequal correlations $r^{(i)}$ with $r(r_i, n_i) \sim 0$	Box M	7.35	32.45	7.45	41.35	7.90	43.85	18.85	24.90
	Schott T1	3.05	30.90	3.15	38.55	3.30	38.75	2.15	14.65
	Schott T2	9.40	31.45	9.45	39.15	11.80	39.60	16.35	15.80
	Schott T3	12.70	30.20	17.35	35.60	22.75	37.65	34.70	15.65
	Levene	7.55	34.90	8.70	14.10	6.10	45.25	7.75	9.20
	R Dispersion	4.25	5.65	3.50	4.15	4.60	5.15	3.70	4.60
unequal correlations $r^{(i)}$ with $r(r_i, n_i) \sim 0.9$	Box M	13.00	82.35	21.95	99.75	12.20	55.45	34.80	99.55
	Schott T1	7.20	81.85	15.45	99.75	14.15	64.65	44.25	99.80
	Schott T2	13.50	81.10	25.30	99.65	25.80	64.25	58.90	99.70
	Schott T3	17.65	79.45	33.65	97.30	36.65	61.70	71.20	96.55
	Levene	2.30	2.95	0.90	0.65	3.75	6.95	1.80	1.10
	R Dispersion	4.70	5.00	4.30	6.10	3.90	4.65	5.25	5.60
unequal correlations $r^{(i)}$ with $r(r_i, n_i) \sim -0.9$	Box M	12.85	90.20	21.15	99.85	10.75	51.45	27.70	98.90
	Schott T1	7.00	85.60	13.65	99.80	4.45	34.20	2.60	94.70
	Schott T2	13.25	84.50	23.50	99.75	14.50	34.75	20.40	93.65
	Schott T3	18.05	83.55	31.40	97.95	25.20	34.95	39.85	87.20
	Levene	6.35	6.35	13.30	16.45	5.40	5.70	8.55	10.90
	R Dispersion	3.85	4.40	3.30	3.35	2.95	3.75	3.40	3.95
varying correlations $r^{(i)}_{j1j2}$ for variables j_1, j_2 within each $\Sigma^{(i)}$	Box M	8.15	34.65	11.40	78.10	7.90	20.05	23.05	75.10
	Schott T1	3.35	34.75	5.60	74.65	6.40	24.70	7.75	71.75
	Schott T2	8.40	34.80	13.50	73.45	14.55	25.40	28.25	70.70
	Schott T3	10.90	33.75	20.05	73.40	25.90	24.45	45.85	70.45
	Levene	3.65	4.00	6.75	34.25	4.40	5.25	8.80	45.95
	R Dispersion	4.25	5.10	3.70	4.30	3.80	4.30	4.25	5.15

Table 4: power rates of an underlying normal distribution at n_i=10, 50 for 6 methods of testing the homogeneity of covariance matrices, for small and large designs, equal und unequal cell counts.

Table 5: Type I error rates of the tests for homogeneity of correlation matrices

	cell counts	equal				unequal			
	design	small		large		small		large	
distribution	method	10	50	10	50	10	50	10	50
normal	Jennrich test	7.80	5.00	12.65	5.85	8.70	6.40	22.05	8.25
	Levene test	5.45	3.65	4.65	3.00	5.95	3.60	7.20	4.25
	Larntz & P	5.45	4.70	7.05	4.80	6.25	5.35	7.70	5.00
	Box M test	10.05	6.70	24.30	9.55	11.90	8.00		10.50
lognormal	Jennrich test	7.45	7.00	16.45	9.95	8.55	8.35	23.50	10.90
	Levene test	5.05	3.25	4.30	2.45	5.10	4.20	6.15	4.20
	Larntz & P	6.40	6.30	8.50	7.65	6.45	6.95	10.65	6.50
	Box M test	10.95	9.10	28.55	15.10	12.45	10.60		16.85
uniform (discrete)	Jennrich test	7.40	6.10	15.05	8.55	9.15	6.4	23.00	10.05
	Levene test	5.65	4.35	5.45	4.45	6.35	4.5	7.60	5.40
	Larntz & P	6.10	5.20	8.25	5.85	7.20	6.2	8.15	5.85
	Box M test	9.90	7.75	27.05	11.95	12.35	8.9		12.60
exponential	Jennrich test	14.40	17.40	36.70	45.60	14.35	16.15	43.0	43.55
	Levene test	4.45	3.65	3.30	3.90	5.55	4.85	7.8	4.30
	Larntz & P	12.30	15.45	22.30	26.15	12.00	14.70	23.0	25.85
	Box M test	21.15	21.40	58.65	54.30	21.55	20.95		53.00
mixed skewed	Jennrich test	8.65	5.65	16.15	8.45	9.55	5.90	24.2	9.50
	Levene test	5.70	4.20	5.60	3.75	6.20	4.45	7.6	4.35
	Larntz & P	6.75	5.30	9.25	5.75	6.55	6.25	10.3	6.15
	Box M test	11.95	7.25	29.05	11.95	12.70	8.25		12.40

Table 5: type I error rates for 4 methods of testing the homogeneity of correlation matrices, for n=10, 50, for 5 distributions, small and large designs, equal und unequal cell counts.

Table 6: Power of the tests for homogeneity of correlation matrices

	cell counts	equal				unequal			
	design	small		large		small		large	
model	method	10	50	10	50	10	50	10	50
unequal correlations $r^{(i)}$ with r(r_i, n_i) ~ 0	Levene test	7.7	36.95	5.60	12.95	11.05	51.5	9.85	22.45
	Larntz & P	14.4	70.15	12.95	52.00	16.20	80.7	15.85	68.20
unequal correlations $r^{(i)}$ with r(r_i, n_i) ~ 0.9	Levene test	8.5	53.5	10.55	70.75	10.60	32.15	35.15	77.65
	Larntz & P	18.0	83.4	39.40	99.80	12.05	53.05	26.30	99.40
unequal correlations $r^{(i)}$ with r(r_i, n_i) ~ - 0.9	Levene test	12.5	72.9	10.95	72.1	7.95	38.05	3.95	52.95
	Larntz & P	21.4	90.5	38.55	99.8	13.95	60.35	31.90	98.90
varying correlations $r^{(i)}_{j1j2}$ for variables j_1,j_2	Levene test	7.25	25.20	6.85	51.55	8.65	36.15	9.7	49.70
	Larntz & P	9.15	43.95	10.25	40.65	10.80	54.95	10.7	41.95

Table 6: power rates of an underlying normal distribution at n=10, 50 for 4 methods of testing the homogeneity of correlation matrices, for small and large designs, equal und unequal cell counts.