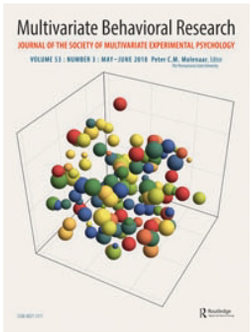

Routledge
Taylor & Francis Group

Multivariate Behavioral Research

ISSN: 0027-3171 (Print) 1532-7906 (Online) Journal homepage: https://www.tandfonline.com/loi/hmbr20

Testing Mean Differences among Groups: Multivariate and Repeated Measures Analysis with Minimal Assumptions

Arne C. Bathke, Sarah Friedrich, Markus Pauly, Frank Konietschke, Wolfgang Staffen, Nicolas Strobl & Yvonne Höller

To cite this article: Arne C. Bathke, Sarah Friedrich, Markus Pauly, Frank Konietschke, Wolfgang Staffen, Nicolas Strobl & Yvonne Höller (2018) Testing Mean Differences among Groups: Multivariate and Repeated Measures Analysis with Minimal Assumptions, Multivariate Behavioral Research, 53:3, 348-359, DOI: 10.1080/00273171.2018.1446320

To link to this article: https://doi.org/10.1080/00273171.2018.1446320

© 2018 The Author(s). Published with license by Taylor & Francis© Arne C. Bathke, Sarah Friedrich, Markus Pauly, Frank Konietschke, Wolfgang Staffen, Nicolas Strobl, and Yvonne Höller

Published online: 22 Mar 2018.

Submit your article to this journal

Article views: 5710

View related articles

View Crossmark data

Citing articles: 17 View citing articles

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=hmbr20

MULTIVARIATE BEHAVIORAL RESEARCH
2018, VOL. 53, NO. 3, 348–359
https://doi.org/10.1080/00273171.2018.1446320

Routledge
Taylor & Francis Group

OPEN ACCESS Check for updates

Testing Mean Differences among Groups: Multivariate and Repeated Measures Analysis with Minimal Assumptions

Arne C. Bathke[a], Sarah Friedrich[b], Markus Pauly[b], Frank Konietschke[c], Wolfgang Staffen[d], Nicolas Strobl[d], and Yvonne Höller[d]

[a]Department of Mathematics, University of Salzburg; Department of Statistics, University of Kentucky; [b]Institute of Statistics, University of Ulm; [c]Department of Mathematical Sciences, University of Texas at Dallas; [d]Department of Neurology, Christian Doppler Medical Centre and Centre for Cognitive Neuroscience, Paracelsus Medical University

ABSTRACT

To date, there is a lack of satisfactory inferential techniques for the analysis of multivariate data in factorial designs, when only minimal assumptions on the data can be made. Presently available methods are limited to very particular study designs or assume either multivariate normality or equal covariance matrices across groups, or they do not allow for an assessment of the interaction effects across within-subjects and between-subjects variables. We propose and methodologically validate a parametric bootstrap approach that does not suffer from any of the above limitations, and thus provides a rather general and comprehensive methodological route to inference for multivariate and repeated measures data. As an example application, we consider data from two different Alzheimer's disease (AD) examination modalities that may be used for precise and early diagnosis, namely, single-photon emission computed tomography (SPECT) and electroencephalogram (EEG). These data violate the assumptions of classical multivariate methods, and indeed classical methods would not have yielded the same conclusions with regards to some of the factors involved.

KEYWORDS

Bootstrap; closed testing; factorial designs; MANOVA; repeated measures

1. Introduction

Almost all interesting data sets are multivariate, that is, they involve more than one variable. Researchers are typically interested in investigating relations, associations, and dependencies between different variables. This is done using several descriptive and inferential techniques. In a more narrow sense, in this manuscript, the term *multivariate analysis* is understood as *statistical inference with several response variables*. However, the methodology described in this manuscript also allows for including several explanatory variables or factors into the model. And the novel tools presented here may also be used in an exploratory fashion, for a descriptive data analysis.

The goal of this manuscript is to provide an approach to inference for means of multivariate data in factorial designs, which does not have to rely on rather limiting assumptions. Traditional analyses of multivariate data such as, for example, those deriving from electroencephalogram (EEG) or single-photon emission computed tomography (SPECT) measurements have often been carried out using essentially univariate techniques (e.g., Moretti, 2015). In such analyses, multivariate responses are either aggregated into one univariate outcome, or separate analyses are performed for different response variables, ideally at least with some adjustment for multiplicity. It is generally a good strategy to perform supplemental marginal analyses, in addition to applying multivariate methods (see also Section 3.3). However, the exclusive use of univariate techniques has in large part been driven by the fact that appropriate inference methods to analyze multivariate data have not existed. Indeed, classical multivariate analysis of variance (MANOVA) techniques (Bartlett, 1939; Dempster, 1958, 1960; Hotelling, 1947, 1951; Lawley, 1938; Nanda, 1950; Pillai, 1955; Wilks, 1946) assume multivariate normal responses with equal covariance matrices across groups, and they are known to perform poorly when covariance matrices do in fact differ and the design is unbalanced (Konietschke, Bathke, Harrar, & Pauly, 2015; Vallejo & Ato, 2012). Unbalancedness and heteroscedasticity are however common real data features that one needs to properly address for valid analyses. For example, in an Alzheimer

CONTACT Arne C. Bathke Arne.Bathke@sbg.ac.at Department of Mathematics, University of Salzburg, 5020 Salzburg, Austria, and Department of Statistics, University of Kentucky, Lexington, Kentucky 40536, USA.

Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/r/HMBR.

Supplemental data for this article can be accessed on the publisher's website.

© 2018 Arne C. Bathke, Sarah Friedrich, Markus Pauly, Frank Konietschke, Wolfgang Staffen, Nicolas Strobl, and Yvonne Höller. Published with license by Taylor & Francis.
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives License (http://creativecommons.org/licenses/by-nc-nd/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited, and is not altered, transformed, or built upon in any way.

data set considered for illustration in this manuscript, empirical variances for different patient groups showed almost 50-fold differences (13.84 vs. 0.28 for variable 6 between AD and SCC, see supplement Table 1).

Only very recently, procedures have been developed that do not suffer from the severe restrictions of classical MANOVA, while at the same time allowing for a factorial design structure of the explanatory variables (Konietschke et al., 2015, see also Pauly, Brunner, & Konietschke, 2015). The present article pursues this resampling-based approach further, in order to develop validated inference methods for a more generalized analysis of not only multivariate, but also repeated measures data. Particularly, we consider the possibility that effects on the response variables may depend on the group levels or other explanatory variables or covariates. In contrast, MANOVA-type methods are typically used to show that groups differ in their effect on a full set of response variables, but these effects are not differentiated any further. However, it is realistic to assume that, for example, two patient groups differ with regard to some responses, while other groups may differ with respect to other response variables. These considerations bear similarities to repeated measures inference, and indeed this aspect of the present manuscript can also be regarded as proposing a new method for the analysis of repeated measures data. Thereby, the underlying model is even rather general: without normality assumption, and without the assumption of covariance matrix equality. The methods proposed in this paper have been implemented in the R-package **MANOVA.RM** (Friedrich, Konietschke, & Pauly, 2016).

Additionally, we identified situations in which multivariate inference should be supplemented by marginal (univariate) methods. In other words, while in most situations, it is advantageous to perform a truly multivariate analysis on multivariate data, there are also cases where marginal analyses add important pieces of information.

Finally, in case of significant effects when using all response variables and all (within-subjects) factors, it is of interest to go a few steps further and find out *which variables* and *which factor levels* are responsible for the significance. In order to accomplish this goal, we make consequent use of the closed testing approach. While closed testing is not new *per se* (see Marcus, Peritz, & Gabriel, 1976), its usefulness for a powerful analysis of multivariate data has not been widely appreciated yet and may deserve more attention.

The new inferential methods proposed in this paper have the potential to be used widely, beyond the neurological applications shown here, since the need for developing and using appropriate multivariate inference methods has already been recognized and articulated across a number of fields of research. For example, in the context of traumatic brain injuries (TBI), where the outcome after TBI is *per definitionem* multidimensional, including neuro-physical disabilities and disturbances in mental functioning, the IMPACT recommendations (Maas et al., 2010) "see a need to explore the feasibility of developing a multidimensional approach to outcome assessment and classification." Bagiella et al. (2010) describe the problem that "no single measure could capture the multidimensional nature of the outcome," and Margulies and Hicks (2009) point out that important deficits could not be identified using univariate functional assessment scales. In other contexts, similar arguments have been made (Vester, 2014). For example, Whitehead, Branson, and Todd (2010) state a "growing interest, especially for trials in stroke, in combining multiple endpoints," while Huang et al. (2009) say that "Parkinson's disease (PD) impairments are multidimensional, making it difficult to choose a single primary outcome."

Upon reviewing the literature on inference methods for multivariate data, there are very few approaches which do not assume at least one of either multivariate normality or covariance matrix equality across groups (or even both). Among these are the permutation-based nonparametric combination methods discussed, for example, in Pesarin and Salmaso (2010) or Pesarin and Salmaso (2012) (see also Anderson, 2001), and the fully nonparametric rank-based tests presented in Bathke and Harrar (2008), Bathke, Harrar, and Madden (2008), Harrar and Bathke (2008a, b), and Liu, Bathke, and Harrar (2011), and implemented in the R package `npmv` (Burchett & Ellis, 2015; Ellis, Burchett, Harrar, & Bathke, 2017). However, these methods are currently limited to the one-way layout, or to some particular factorial design situations (Hahn & Salmaso, 2015, Bathke & Harrar, 2016). Thus, they are not applicable to data from complex factorial designs, such as the AD data described in detail below. Also, methodologically, the articles mentioned are not directly comparable to our approach, as the hypotheses tested are formulated using the distribution functions, or exchangeability of the observation vectors is postulated. In contrast, the methods presented in this article test hypotheses that are formulated using contrasts in terms of mean vectors, and they do not assume exchangeability.

Other procedures based on testing mean vectors, but derived under the assumption of multivariate normality, have been presented for different (one- and two-way) designs, by Nel and Van der Merwe (1986), Krishnamoorthy and Yu (2004, 2012), Belloni and Didier (2008), Girón and Castillo (2010), Krishnamoorthy and Lu (2010), Zhang (2011, 2012, 2013), Xu, Yang, Abula, and Qin (2013), Zhang and Liu (2013), and Kawasaki and Seo (2015).

Without the normality assumption, but requiring homogenous covariance matrices, Van Aelst and Willems (2011) have derived robust one-way MANOVA tests,

which are implemented in the R package FRB (Van Aelst & Willems, 2013).

Apart from Konietschke et al. (2015), the only other mean-based inference method using a multivariate factorial model without normality or equal covariance matrix assumption is that of Harrar and Bathke (2012). However, due to its design limitations, it cannot provide inferential answers to the research questions formulated above regarding the Alzheimer data.

1.1. Alzheimer's disease data example

The demographic development in most Western countries comes along with a rapidly growing incidence of dementia (Barnes & Yaffe, 2011; Prince et al., 2013). Several strategies are being developed to face this challenge, among them early diagnosis, early treatment, and, consequently, prevention of a dementing course (Bateman, 2015). For an accurate and early diagnosis, several examination modalities have been evaluated. For example, SPECT is a well examined and established tool to differentiate Alzheimer's disease (AD) from other forms, such as frontotemporal dementia and dementia with Lewy bodies (Yeo, Lim, Khan, & Pal, 2013). While SPECT is considered to be a cheap diagnostic tool, the costs for an EEG are even lower. The EEG has the additional advantages of being highly available, free of radiation hazards, and noninvasive. It also appears to have considerable diagnostic utility in early-onset dementia (Micanovic & Pal, 2014) via extraction of biomarkers (Vecchio et al., 2013).

Despite its promise, biomarker research faces some basic problems. A typical EEG-based biomarker extraction yields several markers obtained from different electrode positions (typically from 21 to 256 channels), possibly being split into different frequency bands (for an example, see Figure 1). Similarly, quantitative analysis of SPECT data requires the evaluation of perfusion values from many possible brain regions of interest.

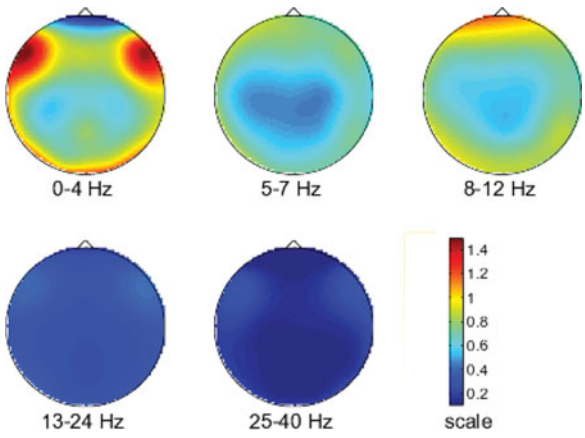

Figure 1. Topographical maps of EEG activity in μV in frequency ranges of interest in a patient sample with AD.

In this article, we consider data from 160 patients who were diagnosed with either AD, mild cognitive impairment (MCI), or subjective cognitive complaints without clinically significant deficits (SCC). The data will be described in more detail in Section 3. Some research questions to be investigated are as follows. Do early forms of AD, namely, subjective cognitive complaints without clinically significant deficits (SCC) and mild cognitive impairment (MCI), differ with regard to average EEG or SPECT feature intensities? Are the differences more pronounced for certain age and sex cohorts? If so, between which of the cognitive impairment stages can the greatest differences be identified? Finally, the structure within the EEG features may exhibit a particular pattern. There may be differences across the regions, modalities, and types of extracted biomarkers (so-called features), or across the spectral distributions. Furthermore, these within-subjects factors may interact with the between-subjects factors disease status, sex, or age.

2. Model

In mathematical terms, multivariate response vectors may be modeled as follows:

$$\mathbf{X}_{ir} = \boldsymbol{\mu}_i + \boldsymbol{\varepsilon}_{ir},\ i = 1, \ldots, d,\ r = 1, \ldots, n_i,\ N = \sum_{i=1}^{d} n_i, \quad (1)$$

where the index i (taking values between 1 and d) represents the treatment group, sample, or, in a factorial design, the treatment combination, while r (between 1 and n_i) denotes the experimental unit or subject on which p-variate observations are being obtained. In order to derive asymptotic results and establish large-sample validity (for large sample sizes n_i) of the proposed methods, the following technical regularity assumptions are necessary:

A. The error terms $\boldsymbol{\varepsilon}_{i1}, \ldots, \boldsymbol{\varepsilon}_{in_i}$ are independent and identically distributed p-dimensional random vectors with $E(\boldsymbol{\varepsilon}_{i1}) = 0$, $\mathrm{Cov}(\boldsymbol{\varepsilon}_{i1}) = \boldsymbol{\Sigma}_i > 0$, and finite fourth moment of their norm, that is, $E(\|\boldsymbol{\varepsilon}_{i1}\|^4) < \infty$, for $i = 1, \ldots, d$; and
B. The different sample sizes n_i grow at the same rate, that is, $n_i/N \to \kappa_i > 0,\ i = 1, \ldots, d$ as $N \to \infty$.

Note that neither normality of the errors nor equality of their variance-covariance matrices $\boldsymbol{\Sigma}_i$ is assumed. The distributions of the error vectors $\boldsymbol{\varepsilon}_{ir}$ may even differ across the groups, as long as their fourth moments are finite. The latter is indeed a mild technical assumption that is needed for the theoretical methodology development. In practice, most data consist of values that are contained between a finite smallest possible value

and a finite largest possible value, so that the technical assumption is quasi always met. However, regarding the small- to moderate-sample performance of the method, it may indeed matter whether data distributions are rather leptokurtic. Several inference procedures then require larger sample sizes in order to perform validly. The simulation results presented in the supplementary materials indicate that the methods proposed here are rather robust in this regard. Also, in simulations by Konietschke et al. (2015, Figure 1), leptokurtic distributions had a larger effect on the power of classical tests without resampling than on tests involving parametric or nonparametric bootstrap.

The vectors from model (1) are aggregated into $\mathbf{X} = (\mathbf{X}'_{11}, \ldots, \mathbf{X}'_{ir}, \ldots, \mathbf{X}'_{dn_d})'$ (length $p \cdot N$), $\boldsymbol{\mu} = (\boldsymbol{\mu}'_1, \ldots, \boldsymbol{\mu}'_i, \ldots, \boldsymbol{\mu}'_d)'$ (length $p \cdot d$), and $\boldsymbol{\varepsilon} = (\boldsymbol{\varepsilon}'_{11}, \ldots, \boldsymbol{\varepsilon}_{ir}, \ldots, \boldsymbol{\varepsilon}'_{dn_d})'$ (length $p \cdot N$), respectively, where $\boldsymbol{\mu}_i = (\mu_i^{(1)}, \ldots, \mu_i^{(p)})'$, $i = 1, \ldots, d$.

2.1. Multivariate hypotheses on between-subjects factors and their interactions

At first sight, the notation introduced above suggests that only a one-way layout is being considered, where the factor levels are indexed by $i = 1, \ldots, d$. However, by splitting up the index i into different sub-indices, factorial structures are introduced within the components of $\boldsymbol{\mu}$ or $\mathbf{X}$. For a complete three-way MANOVA, for example, using the three between-subjects factors age, sex, and diagnosis, as suggested by the Alzheimer data set, the index i is split up into three indices $i = 1, \ldots, a$, $j = 1, \ldots, b$, and $k = 1, \ldots, c$, each corresponding to one of the factors A (sex), B (diagnosis), and C (age group) involved in the study. Then, for example, $\boldsymbol{\mu} = (\boldsymbol{\mu}'_{111}, \ldots, \boldsymbol{\mu}'_{abc})'$, where the entries $\boldsymbol{\mu}_{ijk}$ are lexicographically ordered, and $d = abc$. In classical MANOVA notation, these treatment mean vectors are then decomposed as $\boldsymbol{\mu}_{ijk} = \boldsymbol{\alpha}_i + \boldsymbol{\beta}_j + \boldsymbol{\gamma}_k + (\boldsymbol{\alpha\beta})_{ij} + (\boldsymbol{\alpha\gamma})_{ik} + (\boldsymbol{\beta\gamma})_{jk} + (\boldsymbol{\alpha\beta\gamma})_{ijk}$, with the terms on the right-hand side of the equation symbol satisfying some identifiability constraints.

In matrix notation, hypotheses for each of the different main, simple, and interaction effects can be formulated using an appropriately chosen contrast matrix $\mathbf{T}$. Denote the m-dimensional identity matrix by $\mathbf{I}_m$, the $m \times m$-matrix of ones by $\mathbf{J}_m$, and the so-called centering matrix by $\mathbf{P}_m = \mathbf{I}_m - \frac{1}{m}\mathbf{J}_m$. The hypothesis of no main effect A has the interpretation that for each level $i = 1, \ldots, a$ of the factor A, the mean response is the same, when averaged over the levels of all other factors. Mathematically, this can be formulated as $H_0(A) : \{\overline{\boldsymbol{\mu}}_{1\cdot\cdot} = \cdots = \overline{\boldsymbol{\mu}}_{a\cdot\cdot}\}$, where $\overline{\boldsymbol{\mu}}_{i\cdot\cdot} = \frac{1}{bc}\sum_{j=1}^{b}\sum_{k=1}^{c}\boldsymbol{\mu}_{ijk}$, $i = 1, \ldots, a$, or in classical notation as $\boldsymbol{\alpha}_1 = \cdots = \boldsymbol{\alpha}_a = \mathbf{0}$. Written in Kronecker product notation, this is equivalent to $\{(\mathbf{P}_a \otimes \frac{1}{b}\mathbf{J}_b \otimes \frac{1}{c}\mathbf{J}_c \otimes \mathbf{I}_p)\boldsymbol{\mu} = \mathbf{0}\}$, Thus, it corresponds to the choice $\mathbf{T} = \mathbf{P}_a \otimes \frac{1}{b}\mathbf{J}_b \otimes \frac{1}{c}\mathbf{J}_c \otimes \mathbf{I}_p$. The Kronecker product notation has a few methodological advantages. For one, theoretical properties of test statistics for the hypotheses being considered essentially boil down to the mathematical properties of the matrices $\mathbf{T}$ that are defining these hypotheses. Second, the notation can be directly translated into a statistical programming language, making implementation and simulation straightforward. And, the beauty of the simple Kronecker product notation structure becomes apparent when considering other possible hypotheses. For example, the hypotheses of no main effect B and C are given by $H_0(B) : \{\overline{\boldsymbol{\mu}}_{\cdot 1\cdot} = \cdots = \overline{\boldsymbol{\mu}}_{\cdot b\cdot}\} \Leftrightarrow \boldsymbol{\beta}_1 = \cdots = \boldsymbol{\beta}_b = \mathbf{0} \Leftrightarrow \{(\frac{1}{a}\mathbf{J}_a \otimes \mathbf{P}_b \otimes \frac{1}{c}\mathbf{J}_c \otimes \mathbf{I}_p)\boldsymbol{\mu} = \mathbf{0}\}$, and $H_0(C) : \{\overline{\boldsymbol{\mu}}_{\cdot\cdot 1} = \cdots = \overline{\boldsymbol{\mu}}_{\cdot\cdot c}\} \Leftrightarrow \boldsymbol{\gamma}_1 = \cdots = \boldsymbol{\gamma}_c = \mathbf{0} \Leftrightarrow \{(\frac{1}{a}\mathbf{J}_a \otimes \frac{1}{b}\mathbf{J}_b \otimes \mathbf{P}_c \otimes \mathbf{I}_p)\boldsymbol{\mu} = \mathbf{0}\}$, respectively, with the averages $\overline{\boldsymbol{\mu}}_{\cdot j\cdot}$ and $\overline{\boldsymbol{\mu}}_{\cdot\cdot k}$ defined accordingly. While main effects are defined by Kronecker products involving exactly one $\mathbf{P}$-matrix (centering matrix), two-way interactions are specified using two centering matrices, three-way interactions involve three, and so forth. The hypothesis of, for example, no interaction effect between factors A and B can be written as $H_0(AB) : \{(\mathbf{P}_a \otimes \mathbf{P}_b \otimes \frac{1}{c}\mathbf{J}_c \otimes \mathbf{I}_p)\boldsymbol{\mu} = \mathbf{0}\}$, equivalent to $(\boldsymbol{\alpha\beta})_{11} = \cdots = (\boldsymbol{\alpha\beta})_{ab} = \mathbf{0}$ in classical notation. Other interaction effects are defined analogously by using the centering matrix $\mathbf{P}$ for those factors whose effect (main or interaction) is of interest, and using the averaging matrix $\mathbf{J}$ for all other factors. For multivariate hypotheses, the last component is always the p-dimensional identity matrix, as multivariate hypotheses consider all response variables (endpoints) simultaneously.

2.2. Hypotheses involving within-subjects factors

There are situations where the response variables are commensurate in the sense that comparisons between them are meaningful, for example, measurements on the same subject at different time points or body locations. In that case, formulating and testing hypotheses involving such comparisons may be of interest, in particular when the response vector is structured by one or more within-subjects factors.

The simplest such hypothesis would be that of no averaged, or *marginal* treatment effect, where averaging is over the means of all p response variables. This can be accomplished by replacing the identity matrix $\mathbf{I}_p$ at the last component of the Kronecker product with the averaging matrix $\mathbf{J}_p$ of the same dimension. There is a major difference in interpretation between such a hypothesis formulation, and the multivariate hypotheses introduced in Section 2.1 Namely, *multivariate* equality

of two treatments assumes that the treatment means agree in each response variable, while the *marginal* effects considered here only assume equality of the treatments when averaging across the responses.

By splitting up the index $s = 1, \ldots, p$ that identifies the response variables, the formulation of many more hypotheses, particularly ones involving within-subjects factors, is facilitated. In the Alzheimer data, different EEG values on each subject may be considered commensurate, especially after standardizing each of the response variables. Also, they are structured by the two factors *brain region* (here: temporal, frontal, and central) and *feature* (here: brain rate and complexity). Denote region by the index $r = 1, \ldots, p_r$ and feature by $s = 1, \ldots, p_s$. Then, each of the six possible combinations is uniquely defined by the index pair (r, s), suggesting a natural way to split up the index labeling the responses. Apparent similarity to repeated measures analysis is not coincidental, as indeed the multivariate model (1) presented in this paper can also be interpreted as a repeated measures model if the response variables are commensurate. Even a rather general repeated measures model: without normality assumption, and without the assumption of covariance matrix equality.

For simplicity, assume in the following that, in addition to the two within-subjects factors *brain region* and *feature*, there is only one between-subjects factor present, whose levels are $i = 1, \ldots, a$ (e.g., *diagnosis*). The mean vector $\boldsymbol{\mu}_i = (\mu_i^{(1)}, \ldots, \mu_i^{(p)})'$ then becomes $\boldsymbol{\mu}_i = (\mu_i^{(11)}, \ldots, \mu_i^{(p_r p_s)})'$, $i = 1, \ldots, a$, where the entries are again lexicographically ordered.

With these definitions, it is possible to formulate the corresponding null hypotheses as follows. The order of matrices in the Kronecker product needs to correspond exactly to the way the entries in the vector $\boldsymbol{\mu}$ are sorted. By convention, the between-subjects factors are listed first, followed by the within-subjects factors. The final matrix $\mathbf{I}_p$ used in the previous section needs to be replaced by appropriate choices, corresponding to the within-subjects factors. In the data set considered, these are the two factors *brain region* and *feature*. Therefore, the role of $\mathbf{I}_p$ is taken by a Kronecker product of two matrices whose dimensions are the respective numbers of levels of these two factors. For example, the hypothesis of no main effect of *brain region* is written as $H_0(R) : \{(\frac{1}{a}\mathbf{J}_a \otimes \mathbf{P}_{p_r} \otimes \frac{1}{p_s}\mathbf{J}_{p_s})\boldsymbol{\mu} = \mathbf{0}\}$. The presence of exactly one $\mathbf{P}$-matrix indicates that a main effect is under investigation. Its dimension is p_r, that is, the number of regions. The hypothesis of no two-way interaction between *diagnosis* and *brain region* requires two $\mathbf{P}$-matrices, at the places corresponding to the between-subjects factor *diagnosis*, and the first within-subjects factor *brain region*: $H_0(AR) : \{(\mathbf{P}_a \otimes \mathbf{P}_{p_r} \otimes \frac{1}{p_s}\mathbf{J}_{p_s})\boldsymbol{\mu} = \mathbf{0}\}$.

2.3. Test statistics

As seen in the preceding two subsections, all relevant hypotheses can be written as $H_0 : \mathbf{T}\boldsymbol{\mu} = \mathbf{0}$, with appropriate choices of $\mathbf{T}$. The corresponding Wald-type test statistic (WTS) is defined as

$$Q_N(\mathbf{T}) = N \cdot \overline{\mathbf{X}}_{\cdot}'\mathbf{T}(\mathbf{T}\hat{\mathbf{V}}_N\mathbf{T})^{+}\mathbf{T}\overline{\mathbf{X}}_{\cdot}, \tag{2}$$

where $\overline{\mathbf{X}}_{\cdot} = (\overline{\mathbf{X}}_{1\cdot}', \ldots, \overline{\mathbf{X}}_{d\cdot}')'$, $\overline{\mathbf{X}}_{i\cdot} = \frac{1}{n_i}\sum_{r=1}^{n_i}\mathbf{X}_{ir}$, and

$$\hat{\mathbf{V}}_N = \operatorname{diag}\Big(\frac{N}{n_i}\widehat{\boldsymbol{\Sigma}}_i : 1 \le i \le d\Big),$$
$$\widehat{\boldsymbol{\Sigma}}_i = \frac{1}{n_i - 1}\sum_{r=1}^{n_i}(\mathbf{X}_{ir} - \overline{\mathbf{X}}_{i\cdot})(\mathbf{X}_{ir} - \overline{\mathbf{X}}_{i\cdot})'. \tag{3}$$

Here, $(\cdot)^{+}$ denotes the Moore–Penrose generalized inverse. The generalized inverse needs to be used in lieu of the regular matrix inverse since the latter may not exist, which would make the WTS Q_N invalid. Konietschke et al. (2015) have shown that, under the technical assumptions mentioned above in Section 2, and under $H_0 : \mathbf{T}\boldsymbol{\mu} = \mathbf{0}$, $Q_N(\mathbf{T})$ has asymptotically, as $N \to \infty$, a central χ^2-distribution with degrees of freedom equal to the rank of $\mathbf{T}$. However, they only considered matrices T where the final component is the identity matrix $\mathbf{I}_p$ (see sec. 4 of Konietschke et al., 2015). This leads to the multivariate hypotheses discussed above in Section 2.1 A closer examination of their method of proof reveals that the established asymptotic results remain correct even when $\mathbf{I}_p$ is replaced by other projection matrices, such as those mentioned in Section 2.2 Even the mathematical theory for the asymptotic model-based bootstrap, often referred to as parametric bootstrap, can be transferred to the more general situation without having to impose any additional assumptions. This opens the door to investigating main and interaction effects of within-subjects factors, as well as marginal effects, thus providing a comprehensive toolbox for the analysis of multivariate and repeated measures data with minimal assumptions—neither equal covariance matrices nor multivariate normality of the data are needed for the proposed methods.

2.4. Bootstrap

The idea behind the parametric bootstrap approach originates from an application of the multivariate central limit theorem. In particular, for any $i = 1, \ldots, d$, $\sqrt{n_i}(\overline{\mathbf{X}}_{i\cdot} - \boldsymbol{\mu}_i)$ is asymptotically normal with mean zero and covariance matrix $\boldsymbol{\Sigma}_i$. Thus, for approximation purposes, the original iid observation vectors, $\mathbf{X}_{i1}, \ldots, \mathbf{X}_{in_i}$, can be replaced in the resample by iid parametric bootstrap vectors generated from the estimated limit

Table 1. Number of observations for the different factor level combinations.

Sex	Age	AD	MCI	SCC	Σ
M	<70	2	15	14	31
M	≥70	10	12	6	28
F	<70	9	13	29	51
F	≥70	15	17	18	50
Σ		36	57	67	160

Table 2. Multivariate analysis of EEG data. Factors age (≥70), diagnosis (AD, MCI, SCC), and sex. WTS is the Wald-type statistic approximated by a χ^2-distribution, PBS denotes the asymptotic model-based "parametric" bootstrap.

	Test statistic	df	WTS *p*-value	PBS *p*-value
sex	15.54	6	0.0164	0.0321
age	18.69	6	0.0047	0.0093
sex*age	5.52	6	0.4792	0.5106
sex	12.60	6	0.0498	0.1132
diagnosis	55.16	12	<0.0001	0.0006
sex*diagnosis	9.79	12	0.6344	0.7462
diagnosis	41.81	12	<0.0001	0.0031
age	7.86	6	0.2488	0.3316
diagnosis*age	9.29	12	0.6777	0.7611

distribution. That is, they are replaced by

$$\mathbf{X}^*_{i1}, \ldots, \mathbf{X}^*_{in_i} \overset{\text{iid}}{\sim} N(\mathbf{0}, \widehat{\mathbf{\Sigma}}_i)$$

for each $i = 1, \ldots, d$. Recalculating the Wald-type test statistic in (2) with the variables $\mathbf{X}^*_{i1}, \ldots, \mathbf{X}^*_{in_i}$ yields $Q^*_N(\mathbf{T})$, the parametric bootstrap version of the WTS. The conditional $(1-\alpha)$-quantiles from its distribution, say $c^*(\alpha)$, are then used as critical values, resulting in the bootstrap test $\varphi^*_N = \mathbf{I}\{Q_N(\mathbf{T}) > c^*(\alpha)\}$.

Konietschke et al. (2015) provided simulation results for different multivariate one- and two-factorial designs. The setting in the present article differs somewhat due to the more complex structure, involving within-subjects factors. In order to investigate whether the bootstrap approach also provides satisfactory approximations to the sampling distribution in this design, we have conducted an extensive simulation study. The results, which support the validity of the proposed method for a wide range of design configurations, can be found in the supplementary material to this paper (see supplement Tables 2 and 3). In particular, we have followed Allignol et al. (2011) and performed so-called "empirical simulations" which are particularly adapted to the current data set and may be seen as a case-sensitive justification for the proposed method. Simulations under alternative (see supplement Figure 1) indicate that the powers of both resampling methods, the proposed parametric bootstrap, as well as the nonparametric bootstrap, are very similar. However, in unbalanced designs, the nonparametric bootstrap test shows a quite liberal behavior under null hypothesis. Therefore, inference based on the parametric bootstrap is generally recommended.

3. Data analysis example

We demonstrate the practical usefulness of the proposed multivariate parametric bootstrap method by investigating questions formulated in a neurological study on cognitive impairments. That is, we examine whether mean differences in EEG- or SPECT-features between SCC, MCI, and AD patients can be discovered using the new inferential method (and comparing with the results from existing methods), when the features are considered as multivariate responses. The data set is described in detail in the supplementary materials.

3.1. Design

The three between-subjects factors considered in the data example were *sex* (men vs. women), *diagnosis* (AD vs. MCI vs. SCC), and *age* (<70 vs. ≥70 years). Additionally, we chose the following within-subjects factors structuring the response vectors. For EEG data, we considered the three selected *brain regions*, as well as *feature* (brain rate or complexity). For SPECT data, *brain region* (with six levels) was used as within-subjects factor.

We did not consider *modality*—that is, EEG vs. SPECT values—as another within-subjects factor because these variables are not commensurate, and despite standardization, the method of data acquisition is very different, so that the assessed brain regions cannot even be matched.

Table 3. Multivariate analysis of EEG data using classical methods. Factors age (≥70), diagnosis (AD, MCI, SCC), and sex. PBS denotes the *p*-value from the parametric bootstrap of the WTS.

	Df	Pillai	Approx *F*	num Df	den Df	Pr(>*F*)	PBS *p*-value
sex	1	0.10784	3.0420	6	151	0.007748	0.0321
age	1	0.11469	3.2603	6	151	0.004828	0.0093
sex*age	1	0.03269	0.8505	6	151	0.533001	0.5106
sex	1	0.12615	3.5850	6	149	0.002391	0.1132
diagnosis	2	0.28963	4.2334	12	300	0.0000035	0.0006
sex*diagnosis	2	0.10375	1.3678	12	300	0.180272	0.7462
diagnosis	2	0.24986	3.5691	12	300	0.00005	0.0031
age	1	0.07388	1.9810	6	149	0.07182	0.3316
age*diagnosis	2	0.05879	0.7572	12	300	0.69442	0.7611

Due to the rather small number of patients in some factor level combinations (e.g., only two male patients aged under 70 were diagnosed with AD, see Table 1), we did not consider a layout including all three between-subjects factors, but instead restricted our analyses to layouts with one or two between-subjects factors, as well as one (SPECT) or two (EEG) within-subjects factors.

When using any two of the between-subjects factors, the minimal sample sizes per factor level combination were 28 (*age* and *sex*), 11 (*age* and *diagnosis*), and 12 (*diagnosis* and *sex*), see Table 1. Our simulation studies given in supplement Appendix B indicate that these sample sizes are sufficient to ensure reasonable performance of the proposed parametric bootstrap procedure. Additionally, one-way layouts have been used as basis for post hoc multiple comparison tests regarding the interesting effects. Here, the minimum cell sample sizes were 78 (*age*), 36 (*diagnosis*), and 59 (*sex*).

For comparison, we present both the results of the classical analysis using the Wald-type statistic with a χ^2-approximation, as well as the parametric bootstrap approach described in Section 2.4 with 10,000 bootstrap runs. These analyses were performed using a completely multivariate approach (see also Konietschke et al., 2015) that has been implemented in the R-package **MANOVA.RM**, see supplement Appendix C for details. All analyses were supplemented by a repeated measures analysis, which allows for the formulation of within-subjects effects when the different responses can be considered commensurate. In the EEG case, the six responses were considered sub-structured by *feature* (two levels) and *region* (three levels), whereas in the SPECT case, they were simply considered as six levels of the unstructured factor *brain region*. An important difference between multivariate and repeated measures or marginal approaches is that in the former case, possible effects of the between-subjects factors are considered in each response variable individually, whereas in the latter cases, they are averaged across some or all of the response variables (see also Section 2.2). Both contribute different pieces of information, as demonstrated below.

In general, the p-values provided in the tables are without correction for multiple testing, unless noted specifically. Note, that here and throughout we will regard results as significant if the p-value is smaller than 5%.

Since the focus of the manuscript is not on the specific data, but rather on the methodology, we will not provide a comprehensive data analysis in full detail here. Instead, we will highlight those aspects that we found noteworthy from a methodological point of view. Several more details regarding the analysis of the EEG and SPECT data can be found in the supplement.

3.2. Comparison to classical methods

Simulation results show that the newly proposed method performs rather reliably even in realistic small to moderate sample size scenarios where the traditional MANOVA fails to meet the nominal Type I error rate. Specifically, in our simulations, the proposed parametric bootstrap-based inference procedure always yielded simulated levels of below 7% at nominal 5%-level, while Wilks' Lambda resulted in simulated levels above 20% in several situations (see Tables 2 and 3 in the supplementary materials).

In order to exemplarily compare our results to standard analyses, we additionally considered the following two situations using the Alzheimer data example.

First, we have performed a multivariate inferential analysis of the EEG response data. Here, the newly proposed method was used, as well as the classical MANOVA tests which would serve as arguably the most likely alternative tool that statistics practitioners would resort to.

As described above, only two of the three between-subjects factors were used in each analysis, due to the very small group sizes that would occur if all three between-subjects factors were used simultaneously. The results of the three different multivariate two-way analyses using the new method are shown in Table 2. When interpreting the results, one should keep in mind that using the two factors *age* and *sex*, the sample sizes per group ($n_i \geq 28$) were substantially larger than in the other two possible layouts ($n_i \geq 11$ or 12, respectively, see Table 1 for details).

For a comparison with classical inference methods, the results for Pillai's trace are displayed in Table 3. The other MANOVA methods implemented in R (Wilks' Λ, the Hotelling-Lawley trace, and Roy's largest root) led to similar results in almost all situations. In particular, as the factors *age* and *sex* consisted of only two groups, all four classical methods were equivalent to Hotelling's T^2 in these cases. In the table, the p-value of the new method is also included as "PBS p-value" for direct comparison. In the model including *sex* and *diagnosis*, all classical methods reported a significant effect of *sex*, which was not supported by the bootstrap p-value. Roy's largest root even resulted in a significant interaction effect in this case (results not shown).

When checking the assumptions of the classical MANOVA procedures, we found that the data were non-normally distributed (based on a multivariate Shapiro–Wilk test, p-value < 0.0001). Furthermore, adjusted quantile plots (using the function `aq.plot` from the package **mvoutlier**) suggested several outliers in the data. Due to the nonnormally distributed data, we could not use Box' M-test to check for homoscedasticity of the covariance matrices, but univariate Levene tests applied to the different response variables suggested variance

Table 4. Three-way layout for SPECT data. Between-subjects factors sex and age. Within-subjects factor brain region. WTS is the Wald-type statistic approximated by a χ^2-distribution, PBS denotes the asymptotic model-based "parametric" bootstrap.

	Test statistic	df	WTS *p*-value	PBS *p*-value
sex	0.01	1	0.9258	0.9262
age	8.22	1	0.0042	0.0061
region	507.27	5	<0.0001	<0.0001
sex*age	0.03	1	0.8671	0.8625
sex*region	16.18	5	0.0063	0.0089
age*region	14.46	5	0.0129	0.0177
sex*age*region	8.12	5	0.1497	0.1773

heterogeneity. That is, all complexity values led to significant results (p-values between < 0.0001 and 0.004) regarding variance heterogeneity. These analyses strongly indicated that the assumptions of classical MANOVA methods, namely, multivariate normality and covariance homogeneity, were violated in the AD data example, thus resulting in possibly misleading conclusions (inflated Type I error) when using standard MANOVA methods.

As a second illustration, we analyzed the SPECT data with regard to the factors *sex*, *age*, and *brain region*. Results from the new method are displayed in Table 4, while those from classical repeated measures ANOVA are shown in Table 5. In contrast to the new parametric bootstrap procedure, the classical ANOVA F-Test did not discover significant interactions between *sex* and *region*, nor between *age* and *region*.

The Shapiro–Wilk test for normality rejected the null hypothesis with p-value < 0.0001. Furthermore, based on a Levene test (p-value $= 0.0445$), we also found evidence against the assumption of equal variances in the groups. Thus, the assumptions of a classical ANOVA were violated, and the results based on the ANOVA F-Test might not have sufficient power to detect effects present in the data.

These examples were chosen to illustrate that the use of classical methods could lead to inflated Type I errors, but also to low power, when their assumptions are violated. We think that the method proposed in the present manuscript features some desirable robustness properties, with regard to these assumptions.

Table 6. Multivariate analysis of SPECT data. Factors age (≥70), diagnosis (AD, MCI, SCC), and sex. WTS is the Wald-type statistic approximated by a χ^2-distribution, PBS denotes the asymptotic model-based "parametric" bootstrap.

	Test statistic	df	WTS *p*-value	PBS *p*-value
sex	17.24	6	0.0084	0.0127
age	21.65	6	0.0014	0.0042
sex*age	10.81	6	0.0944	0.1176
sex	14.70	6	0.0227	0.0455
diagnosis	61.55	12	<0.0001	<0.0001
sex*diagnosis	5.73	12	0.9292	0.9517
diagnosis	59.53	12	<0.0001	0.0001
age	16.36	6	0.0120	0.0258
diagnosis*age	11.01	12	0.5281	0.6419

However, we would also like to caution from an overoptimistic or naïve use of the new methods, as illustrated in the next two examples.

3.3. Multivariate vs. marginal or repeated measures analysis?

For SPECT, we chose six relevant response variables in order to have a fair comparison with the EEG analysis, which also used a six-dimensional response. Results from multivariate two-way inference are shown in Table 6. Here, the multivariate effects were significant for each of the between-subjects factors *diagnosis*, *age*, and *sex*, but for none of their pairwise interactions.

Contrary to the EEG analysis, the six variables considered here are not structured factorially. Instead, in a repeated measures type analysis, we may simply regard them as levels of a within-subjects factor *brain region*. Together with using two of the three between-subjects factors *age*, *sex*, and *diagnosis* at a time, we obtain different three-way layouts. The results for the layout involving *sex* and *diagnosis*, as well as *brain region*, are shown in Table 7. Those for the other two configurations can be found in the supplement Tables 12 and 13, respectively.

A comparison of multivariate and marginal SPECT analyses with regard to the factor *sex*, which is discovered as significant in the multivariate analysis ($p = 0.0455$), but rather nonsignificant in the repeated measures analysis ($p = 0.9231$), reveals an important advantage of

Table 5. Three-way layout for SPECT data. Between-subjects factors sex and age. Within-subjects factor brain region. PBS denotes the asymptotic model-based "parametric" bootstrap, for comparison.

	Df	Sum Sq	Mean Sq	*F* value	Pr(>*F*)	PBS *p*-value
sex	1	0.0126	0.0126	0.0063	0.9370	0.9262
age	1	73.3604	73.3604	36.4026	0.0000	0.0061
region	5	190.1318	38.0264	18.8693	0.0000	<0.0001
sex*age	1	0.2257	0.2257	0.1120	0.7380	0.8625
sex*region	5	14.2690	2.8538	1.4161	0.2158	0.0089
age*region	5	6.4003	1.2801	0.6352	0.6729	0.0177
sex*age*region	5	7.1688	1.4338	0.7115	0.6149	0.1773

Table 7. Three-way layout for SPECT data. Between-subjects factors sex and diagnosis. Within-subjects factor brain region. WTS is the Wald-type statistic approximated by a χ^2-distribution, PBS denotes the asymptotic model-based "parametric" bootstrap.

	Test statistic	df	WTS *p*-value	PBS *p*-value
sex	0.01	1	0.9246	0.9231
diagnosis	51.23	2	<0.0001	<0.0001
region	426.56	5	<0.0001	<0.0001
sex*diagnosis	0.91	2	0.6333	0.6374
sex*region	14.16	5	0.0146	0.0264
diagnosis*region	18.31	10	0.0500	0.1119
sex*diagnosis*region	5.37	10	0.8651	0.8936

the truly multivariate approach. In the marginal and repeated measures analyses, effects are averaged across the response variables, whereas the multivariate analysis considers effect contributions of each of the responses individually, while taking their correlation into account by construction of the test statistic.

In this case, the effects of the individual SPECT response variables were in part small and would not lead to significance using classical variable-wise univariate approaches (supplement Tables 15 and 16). However, each response added information, and only the multivariate analysis was able to take advantage of this information. In this case, it did not make sense to average across the SPECT responses, as this led to the masking of some of the available information (see Tables 6 and 7, as well as supplement Tables 12 and 13).

The SPECT analyses demonstrate that considering several variables in a truly multivariate fashion together may provide more information than many individual univariate analyses. Note that it is not necessary for the method that the effect directions match for the different variables.

Regarding the EEG data, the results of multivariate two-way analyses were shown in Table 2. In the corresponding repeated measures analyses incorporating between- and within-subjects factors (as described in Section 2.2), the resulting designs are each four-way layouts using within-subjects factors *brain region* (frontal, central, temporal) and *feature* (brain rate, complexity), as well as two of the between-subjects factors *age*, *diagnosis*, and *sex*. Results using the two between-subjects factors *diagnosis* and *sex* are shown in Table 8. Those for the other two choices of between-subjects factor pairs (*diagnosis* and *age*, *sex* and *age*) are given in supplement Tables 6 and 7.

When comparing the results from Tables 2 (multivariate analysis) and 8 (marginal and repeated measures analysis), one notices their agreement on the significance of diagnosis, and on the lack of an interaction effect between diagnosis and sex. However, there was also a notable difference in the form of a significant ($p = 0.0048$) marginal effect of sex, while the multivariate effect of sex was not

Table 8. Marginal effects/repeated measures analysis. Four-way layouts for EEG data. Between-subjects factors sex and diagnosis (AD, MCI, SCC). Within-subjects factors brain region (frontal, central, temporal) and feature (brain rate, complexity). WTS stands for the classical Wald-type statistic approximated by a χ^2-distribution, whereas PBS denotes the asymptotic model-based "parametric" bootstrap procedure.

Effect	Test statistic	df	WTS *p*-value	PBS *p*-value
sex	9.97	1	0.0016	0.0048
diagnosis	42.38	2	<0.0001	<0.0001
feature	0.09	1	0.7687	0.7728
region	0.07	2	0.9658	0.9662
sex*diagnosis	3.78	2	0.1513	0.1742
sex*feature	2.17	1	0.1410	0.1524
sex*region	0.88	2	0.6454	0.6615
diagnosis*feature	5.32	2	0.0701	0.0913
diagnosis*region	6.12	4	0.1903	0.2316
feature*region	0.65	2	0.7216	0.7461
sex*diagnosis*feature	1.74	2	0.4199	0.4347
sex*diagnosis*region	1.53	4	0.8210	0.8396
sex*feature*region	0.42	2	0.8095	0.8194
diagnosis*feature*region	7.14	4	0.1286	0.1788
sex*diagnosis*feature*region	2.27	4	0.6855	0.7109

significant ($p = 0.1132$). This may happen when data are analyzed as multivariate p-dimensional, but the responses are highly correlated, so that a lower-dimensional set of $q < p$ variables carries all the relevant information. In our case, empirical absolute correlations between the EEG variables ranged from 0.9266 to 0.9991 for men, and between 0.8894 and 0.9987 for women. This contradicts the implicit assumption of a multivariate approach that each variable contributes useful information, reflected by the degrees of freedom, which equaled six for the multivariate test, while there was only one degree of freedom in the marginal analysis of the effect of sex. We emphasize this point here as a caveat regarding a naïve use of multivariate inference methods in general. They may suffer when the "relevant" response space has smaller dimension than p. Investigating the correlation structure between responses is always advisable, and an additional marginal analysis may be useful, where appropriate.

So, when trying to decide between multivariate or marginal/repeated measures analysis—one may consider doing both and looking carefully at the findings.

3.4. Summarizing the general findings from the data set

The results shown here and in the supplement indicate the following. EEG features differed on average between SCC and MCI, and between SCC and AD. The two assessed EEG features seemed to be robust against normal aging effects, and consistent with each other. Temporal and frontal regions may play a relevant role in aging. SPECT perfusion values differed between AD and the other

patient groups, so that EEG and SPECT perfectly complemented one another. In most regions, perfusion was affected by age, which might reflect normal aging processes. We did not find any interaction effects between age, sex, and diagnosis, using EEG or SPECT data.

There were significant effects of sex in this clinical sample. Generally, healthy women have shown a higher amplitude than healthy men in the resting EEG (Wada, Takizawa, Jiang, & Yamaguchi, 1994), and higher coherence values, especially for interhemispheric connections in the delta, theta, and beta range (Wada et al., 1996). Our findings also suggest that altered brain patterns in the demented population should be examined in detail for sex differences.

Main findings of the data analysis have confirmed the conjectured effects that early-onset dementia has on EEG, and there were no interactions of diagnosis with any of the demographic between-subjects factors age and sex. EEG is a cheap diagnostic and noninvasive tool. Its diagnostic utility appears to remain stable across different age and sex cohorts, although as a limitation to this conclusion, it should be mentioned that the study population consisted of mostly elderly people, about half of them 70 years and older.

4. Discussion

Until recently, no valid methodology had been developed for the inferential analysis of multivariate data from factorial designs, unless equal covariance matrices across groups, or multivariate normality could be assumed. For realistic data applications, typically neither of these assumptions is reasonable, as was also apparent for the EEG and SPECT data considered. Here, an analysis based on classical MANOVA or repeated measures ANOVA techniques produced p-values that were most likely too small (corresponding to a Type I error inflation), but also p-values that were most likely too large (corresponding to low power). The assumptions of covariance matrix homoscedasticity and multivariate normality were clearly violated, which may have drastic effects on the classical tests.

The methodology pursued here is based on an asymptotic model-based "parametric" bootstrap approach with rather general asymptotic validity and good finite sample performance (see Konietschke et al., 2015). In addition to demonstrating the usefulness of this approach in real data analysis, and the possible insights to be gained, we have extended the methodology by enabling inference not only for between-subjects factors, as is common for multivariate inference, but also for within-subjects factors and the interactions of all factors involved. This corresponds to a repeated measures approach or profile analysis. Where such an analysis is applicable, it substantially extends the scope of the possible inferential techniques. In the example data, such an extension was sensible due to the commensurate nature of the responses within the respective groups of EEG and SPECT variables. Resulting marginal effects analyses were less influenced by multicollinearity of the responses than a multivariate approach, and thus they may provide useful additional information.

We hope to have demonstrated the potential of novel resampling-based multivariate and marginal or repeated measures methods for factorial designs when data do not follow classical assumptions.

Supplementary Materials

In a supplementary file, we describe the discussed data example in detail, and provide simulation results for a design adopted from the discussed data example. Moreover, additional tables with empirical covariance matrices and several results tables from data analyses with different design configurations are included. In addition, the R code for the evaluation of the data set is presented, along with some more details on data extraction.

Article information

Conflict of Interest Disclosures: Each author signed a form for disclosure of potential conflicts of interest. No authors reported any financial or other conflicts of interest in relation to the work described. The only financial support received was from entities that were not influencing the research and findings described in the manuscript in any way.

Ethical Principles: The authors affirm having followed professional ethical guidelines in preparing this work. These guidelines include obtaining informed consent from human participants, maintaining ethical treatment and respect for the rights of human or animal participants, and ensuring the privacy of participants and their data, such as ensuring that individual participants cannot be identified in reported results or from publicly available original or archival data.

Funding: This work was supported by Grants I 2697-N31 and KLI 12-B00 from the Austrian Science Fund (FWF) and by Grant DFG-PA 2409/3-1 from the German Research Foundation (DFG).

Role of the Funders/Sponsors: None of the funders or sponsors of this research had any role in the design and conduct of the study; collection, management, analysis, and interpretation of data; preparation, review, or approval of the manuscript; or decision to submit the manuscript for publication.

References

Allignol, A., Schumacher, M., Wanner, C., Drechsler, C., & Beyersmann, J. (2011). Understanding competing risks: A simulation point of view. *BMC Medical Research Methodology*, *11*(1), 1. doi:10.1186/1471-2288-11-86

Anderson, M. (2001). A new method for non-parametric multivariate analysis of variance. *Austral Ecology*, *26*(1), 32–46. doi:10.1111/j.1442-9993.2001.01070.pp.x

Bagiella, E., Novack, T. A., Ansel, B., Diaz-Arrastia, R., Dikmen, S., Hart, T., & Temkin, N. (2010). Measuring outcome in traumatic brain injury treatment trials: Recommendations from the traumatic brain injury clinical trials network. *The Journal of Head Trauma Rehabilitation*, *25*(5), 375–382. doi:10.1097/HTR.0b013e3181d27fe3

Barnes, D., & Yaffe, K. (2011). The projected effect of risk factor reduction on Alzheimer's disease prevalence. *Lancet Neurology*, *10*, 819–828. doi:10.1016/S1474-4422(11)70072-2

Bartlett, M. S. (1939). A note on tests of significance in multivariate analysis. In *Mathematical proceedings of the Cambridge Philosophical Society* (Vol. *35*, pp. 180–185). Cambridge: Cambridge University Press. doi:10.1017/S0305004100020880.

Bateman, R. (2015). Alzheimer's disease and other dementias: Advances in 2014. *Lancet Neurology*, *14*, 4–6. doi:10.1016/S1474-4422(14)70301-1

Bathke, A. C., & Harrar, S. W. (2008). Nonparametric methods in multivariate factorial designs for large number of factor levels. *Journal of Statistical Planning and Inference*, *138*(3), 588–610. doi:10.1016/j.jspi.2006.11.004

Bathke, A. C., & Harrar, S. W. (2016). Rank-based inference for multivariate data in factorial designs. In Regina Y. Liu, & Joseph W. McKean (Eds.), *Robust rank-based and nonparametric methods* (pp. 121–139). New York: Springer.

Bathke, A. C., Harrar, S. W., & Madden, L. V. (2008). How to compare small multivariate samples using nonparametric tests. *Computational Statistics and Data Analysis*, *52*(11), 4951–4965. doi:10.1016/j.csda.2008.04.006

Belloni, A., & Didier, G. (2008). On the Behrens–Fisher problem: A globally convergent algorithm and a finite-sample study of the Wald, LR and LM tests. *The Annals of Statistics*, *36*(5), 2377–2408. doi:10.1214/07-AOS528.

Burchett, W. W., & Ellis, A. R. (2015). *npmv* (R Package Version 2.3). Retrieved from http://CRAN.R-project.org/package=npmv

Dempster, A. P. (1958). A high dimensional two sample significance test. *The Annals of Mathematical Statistics*, 995–1010. Retrieved from http://www.jstor.org/stable/2236942.

Dempster, A. P. (1960). A significance test for the separation of two highly multivariate small samples. *Biometrics*, *16*(1), 41–50. doi:10.2307/2527954

Ellis, A. R., Burchett, W. W., Harrar, S. W., & Bathke, A. C. (2017). Nonparametric inference for multivariate data: The r package npmv. *Journal of Statistical Software*, *76*(4), 1–18. doi:10.18637/jss.v076.i04

Friedrich, S., Konietschke, F., & Pauly, M. (2016). MANOVA.RM: Analysis of Multivariate Data and Repeated Measures Designs (R Package Version 0.0.4). Retrieved from https://CRAN.R-project.org/package=MANOVA.RM.

Girón, F., & del Castillo, C. (2010). The multivariate Behrens–Fisher distribution. *Journal of Multivariate Analysis*, *101*(9), 2091–2102. doi:10.1016/j.jmva.2010.04.008

Hahn, S., & Salmaso, L. (2015). A comparison of different synchronized permutation approaches to testing effects in two-level two-factor unbalanced ANOVA designs. *Statistical Papers*, *58*(123), 123–146. doi:10.1007/s00362-015-0690-2.

Harrar, S., & Bathke, A. (2012). A modified two-factor multivariate analysis of variance: Asymptotics and small sample approximations (and erratum). *Annals of the Institute of Statistical Mathematics*, *64*(1 & 5), pp. 135–165 & 1087. doi:10.1007/s10463-010-0299-0; erratum at doi:10.1007/s10463-012-0364-y.

Harrar, S. W., & Bathke, A. C. (2008). A nonparametric version of the Bartlett-Nanda-Pillai multivariate test. Asymptotics, approximations, and applications. *American Journal of Mathematical and Management Sciences*, *28*(3–4), 309–335. doi:10.1080/01966324.2008.10737731

Hotelling, H. (1947). Multivariate quality control—illustrated by the air testing of sample bombsights. In Eisenhart C., Hastay M. W., & Wallis W. A. (Eds.), *Techniques of statistical analysis* (pp. 111–184). New York: McGraw-Hill.

Hotelling, H. (1951). *A generalized t test and measure of multivariate dispersion*. Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability, pp. 23–41. Retrieved from http://projecteuclid.org/euclid.bsmsp/1200500217.

Huang, P., Goetz, C. G., Woolson, R. F., Tilley, B., Kerr, D., Palesch, Y., Elm, J., Ravina, B., Bergmann, K. J., & Kieburtz, K. (2009). Using global statistical tests in long-term Parkinson's disease clinical trials. *Movement Disorders*, *24*(12), 1732–1739. doi:10.1002/mds.22645

Kawasaki, T., & Seo, T. (2015). A two sample test for mean vectors with unequal covariance matrices. *Communications in Statistics-Simulation and Computation*, *44*(7), 1850–1866. doi:10.1080/03610918.2013.824587

Konietschke, F., Bathke, A., Harrar, S., & Pauly, M. (2015). Parametric and nonparametric bootstrap methods for general MANOVA. *Journal of Multivariate Analysis*, *140*, 291–301. doi:10.1016/j.jmva.2015.05.001

Krishnamoorthy, K., & Lu, F. (2010). A parametric bootstrap solution to the MANOVA under heteroscedasticity. *Journal of Statistical Computation and Simulation*, *80*(8), 873–887. doi:10.1080/00949650902822564

Krishnamoorthy, K., & Yu, J. (2004). Modified Nel and Van der Merwe test for the multivariate Behrens–Fisher problem. *Statistics & Probability Letters*, *66*(2), 161–169. doi:10.1016/j.spl.2003.10.012

Krishnamoorthy, K., & Yu, J. (2012). Multivariate Behrens–Fisher problem with missing data. *Journal of Multivariate Analysis*, *105*(1), 141–150. doi:10.1016/j.jmva.2011.08.019

Lawley, D. (1938). A generalization of Fisher's z test. *Biometrika*, *30*(1–2), 180–187. doi:10.2307/2332232

Liu, C., Bathke, A., & Harrar, S. (2011). A nonparametric version of Wilks' lambda – asymptotic results and small sample approximations. *Statistics and Probability Letters*, *81*, 1502–1506. doi:10.1016/j.spl.2011.04.012

Maas, A. I., Steyerberg, E. W., Marmarou, A., McHugh, G. S., Lingsma, H. F., Butcher, I., Lu, J., Weir, J., Roozenbeek, B., & Murray, G. D. (2010). IMPACT recommendations for improving the design and analysis of clinical trials in moderate to severe traumatic brain injury. *Neurotherapeutics*, *7*(1), 127–134. doi:10.1016/j.nurt.2009.10.020

Marcus, R., Peritz, E., & Gabriel, K. (1976). On closed test procedures with special reference to ordered analysis of variance. *Biometrika*, *63*(3), 655–660. doi:10.2307/2335748

Margulies, S., & Hicks, R. (2009). Combination therapies for traumatic brain injury: Prospective considerations. *Journal of Neurotrauma*, *26*(6), 925–939. doi:10.1089/neu.2008-0794

Micanovic, C., & Pal, S. (2014). The diagnostic utility of EEG in early-onset dementia: A systematic review of the literature with narrative analysis. *Journal of Neural Transmission*, *121*, 59–69. doi:10.1007/s00702-013-1070-5

Moretti, D. (2015). Theta and alpha EEG frequency interplay in subjects with mild cognitive impairment: Evidence from EEG, MRI, and SPECT brain modifications. *Frontiers in Aging Neuroscience*, *7*, 31. doi:10.3389/fnagi.2015.00031.

Nanda, D. (1950). Distribution of the sum of roots of a determinantal equation under a certain condition. *The Annals of Mathematical Statistics*, *21*(3), 432–439. Retrieved from http://www.jstor.org/stable/2236498.

Nel, D., & Van der Merwe, C. (1986). A solution to the multivariate Behrens-Fisher problem. *Communications in Statistics-Theory and Methods*, *15*(12), 3719–3735. doi:10.1080/03610928608829342

Pauly, M., Brunner, E., & Konietschke, F. (2015). Asymptotic permutation tests in general factorial designs. *Journal of the Royal Statistical Society: Series B*, *77*(2), 461–473. doi:10.1111/rssb.12073

Pesarin, F., & Salmaso, L. (2010). *Permutation tests for complex data: Theory, applications and software*. New York: John Wiley & Sons.

Pesarin, F., & Salmaso, L. (2012). A review and some new results on permutation testing for multivariate problems. *Statistics and Computing*, *22*(2), 639–646. doi:10.1007/s11222-011-9261-0

Pillai, K. (1955). Some new test criteria in multivariate analysis. *The Annals of Mathematical Statistics*, *26*(1), 117–121. Retrieved from https://projecteuclid.org/euclid.aoms/1177728599.

Prince, M., Bryce, R., Albanese, E., Wimo, A., Ribeiro, W., & Ferri, C. (2013). The global prevalence of dementia: A systematic review and metaanalysis. *Alzheimers Dement*, *9*, 63–75. doi:10.1016/j.jalz.2012.11.007

Vallejo, G., & Ato, M. (2012). Robust tests for multivariate factorial designs under heteroscedasticity. *Behavior Research Methods*, *44*(2), 471–489. doi:10.3758/s13428-011-0152-2

Van Aelst, S., & Willems, G. (2011). Robust and efficient one-way MANOVA tests. *Journal of the American Statistical Association*, *106*(494), 706–718. doi:10.1198/jasa.2011.tm09748

Van Aelst, S., & Willems, G. (2013). Fast and robust bootstrap for multivariate inference: The R package FRB. *Journal of Statistical Software*, *53*(3), 1–32. doi:10.18637/jss.v053.i03

Vecchio, F., Babiloni, C., Lizio, R., Fallani Fde, V., Blinowska, K., Verrienti, G., Frisoni, G., & Rossini, P. (2013). Resting state cortical EEG rhythms in Alzheimer's disease: Toward EEG markers for clinical applications: A review. *Supplements to Clinical Neurophysiology*, *62*, 223–236. doi:10.1016/B978-0-7020-5307-8.00015-6

Vester, J. (2014). *Multivariate inference methods – a new start in neurosciences clinical research*. Presentation at Workshop Multivariate Inference Methods with Applications, International Biometric Society, German Region, Düsseldorf.

Wada, Y., Nanbu, Y., Kadoshima, R., Jiang, Z., Koshino, Y., & Hashimoto, T. (1996). Interhemispheric EEG coherence during photic stimulation: Sex differences in normal young adults. *International Journal of Psychophysiology*, *22*, 45–51. doi:10.1016/0167-8760(96)00011-6

Wada, Y., Takizawa, Y., Jiang, Z., & Yamaguchi, N. (1994). Gender differences in quantitative EEG at rest and during photic stimulation in normal young adults. *Clinical Electroencephalography*, *25*, 81–85. doi:10.1177/155005949402500209

Whitehead, J., Branson, M., & Todd, S. (2010). A combined score test for binary and ordinal endpoints from clinical trials. *Statistics in Medicine*, *29*(5), 521–532. doi:10.1002/sim.3822

Wilks, S. S. (1946). Sample criteria for testing equality of means, equality of variances, and equality of covariances in a normal multivariate distribution. *The Annals of Mathematical Statistics*, *17*(3), 257–281. Retrieved from http://www.jstor.org/stable/2236125.

Xu, L.-W., Yang, F.-Q., Abula, A., & Qin, S. (2013). A parametric bootstrap approach for two-way ANOVA in presence of possible interactions with unequal variances. *Journal of Multivariate Analysis*, *115*, 172–180. doi:10.1016/j.jmva.2012.10.008

Yeo, J., Lim, X., Khan, Z., & Pal, S. (2013). Systematic review of the diagnostic utility of SPECT imaging in dementia. *European Archives of Psychiatry and Clinical Neuroscience*, *263*, 539–552. doi:10.1007/s00406-013-0426-z

Zhang, J.-T. (2011). Two-way MANOVA with unequal cell sizes and unequal cell covariance matrices. *Technometrics*, *53*(4), 426–439. doi:10.1198/TECH.2011.10128

Zhang, J.-T. (2012). An approximate Hotelling T^2-test for heteroscedastic one-way MANOVA. *Open Journal of Statistics*, *2*, 1. doi:10.4236/ojs.2012.21001

Zhang, J.-T. (2013). Tests of linear hypotheses in the ANOVA under heteroscedasticity. *International Journal of Advanced Statistics and Probability*, *1*(2), 9–24. doi:10.14419/ijasp.v1i2.908

Zhang, J.-T., & Liu, X. (2013). A modified Bartlett test for heteroscedastic one-way MANOVA. *Metrika*, *76*(1), 135–152. doi:10.1007/s00184-011-0379-z