

---

# Resampling-Based Inference Methods for Repeated Measures Data with Missing Values

---

DISSERTATION

in partial fulfillment of the requirements for the degree of  
*Doktor der Naturwissenschaften*  
submitted to the

Department of Statistics  
TU Dortmund University

by

**Lubna Amro**

January, 2022

Primary referee: Prof. Dr. Markus Pauly  
Secondary referee: Prof. Dr. Katja Ickstadt  
Commission chairperson: Prof. Dr. Claus Weihs  
Assessor: Dr. Uwe Ligges  
Date of the oral examination: 23.03.2022



# Acknowledgments

First and foremost, I would like to express my deepest and sincerest gratitude to my supervisor, Prof. Markus Pauly, for his insightful guidance, consistent support, and encouragement throughout the conduct of this research project. His intellectual bravery, enthusiasm for research, and admirable work ethic are inspiring.

Besides my supervisor, I would like to thank Prof. Katja Ickstadt for agreeing to be the second supervisor of my dissertation. Further, I would like to thank Prof. Claus Weihs and Dr. Uwe Ligges for accepting to be on my thesis committee.

I would also like to thank my co-author Prof. Frank Konietschke for the insightful comments and suggestions that substantially contributed to this work. In addition, I would like to thank all my colleagues for good company and interesting discussions throughout the last few years.

Furthermore, I gratefully acknowledge the German Academic Exchange Service (DAAD) for granting me a scholarship and for giving me the opportunity to pursue my doctoral degree in Germany.

Finally, I would like to thank my family for their encouragement, unconditional love, and support, without which I would not be where I am today. Thank you for always being there for me. I hope to build a professional life that you will be proud of.

*"Life shrinks or expands in proportion to one's courage."  
-Anais Nin*

# Abstract

The primary objective of this dissertation was to (i) develop novel resampling approaches for handling repeated measures data with missing values, (ii) compare their empirical power against other existing approaches using a Monte Carlo simulation study, and (iii) pinpoint the limitations of some common approaches, particularly for small sample sizes. This dissertation investigates four different statistical problems. The first is semiparametric inference for comparing means of matched pairs with missing data in both arms. Therein, we propose two novel randomization techniques; a weighted combination test and a multiplication combination test. They are based upon combining separate results of the permutation versions of the paired t-test and Welch test for the completely observed pairs and the incompletely observed components, respectively. As second problem, we consider the same setting but missingness in one arm only. There, we investigate a Wald-type statistic (WTS), an ANOVA-type statistic (ATS), and a modified ANOVA-type statistic (MATS). However, ATS and MATS are not distribution free under the null hypothesis, and WTS suffers from the slow convergence to its limiting  $\chi^2$  distribution. Thus, we develop asymptotic model-based bootstrap versions of these tests. The third problem is on nonparametric rank-based inference for matched pairs with incompleteness in both arms. In this more general setup, the only requirement is that the marginal distributions are not one point distributions. Therein, we propose novel multiplication combination tests that can handle three different testing problems, including the nonparametric Behrens-Fisher problem ( $H_0^p : \{p = 1/2\}$ ). Finally, the fourth problem is nonparametric rank-based inference for incompletely observed factorial designs with repeated measures. Therein, we develop a wild bootstrap approach combined with quadratic form-type test statistics (WTS, ATS, and MATS). These rank-based methods can be applied to both continuous and ordinal or ordered categorical data and (some) allow for singular covariance matrices. In addition to theoretically proving the asymptotic correctness of all the proposed procedures, extensive simulation studies demonstrate their favorable small samples properties in comparison to classical parametric tests. We also motivate and validate our approaches using real-life data examples from a variety of fields.

# Contents

## List of Contributed Articles

## Abbreviations

<b>I</b>	<b>Introduction</b>	<b>1</b>
<b>1</b>	<b>Motivation</b>	<b>3</b>
<b>2</b>	<b>Missing Data</b>	<b>8</b>
<b>3</b>	<b>Statistical Models</b>	<b>10</b>
3.1	Semiparametric Models . . . . .	10
3.1.1	Matched Pairs with Missing Values in Both Arms . . . . .	10
3.1.2	Matched Pairs with Missing Values in a Single Arm . . . . .	11
3.2	Nonparametric Models . . . . .	12
3.2.1	Matched Pairs with Missing Values in Both Arms . . . . .	12
3.2.2	Repeated Measures with Missing Values . . . . .	13
<b>4</b>	<b>Resampling Methods</b>	<b>15</b>
4.1	Permutation Test . . . . .	16
4.2	Bootstrap . . . . .	17
4.2.1	Parametric Bootstrap . . . . .	17
4.2.2	Wild Bootstrap . . . . .	18
<b>5</b>	<b>Summary of the Scientific Articles</b>	<b>19</b>
5.1	Article I: Permuting Incomplete Paired Data: a novel exact and asymptotic correct randomization test. . . . .	19
5.2	Article II: Multiplication-Combination Tests for Incomplete Paired Data. . . . .	21
5.3	Article III: Asymptotic Based Bootstrap Approach for Matched Pairs with Missingness in a Single-arm. . . . .	23
5.4	Article IV: Incompletely Observed Nonparametric Factorial Designs with Repeated Measurements: A wild bootstrap approach. . . . .	25
<b>6</b>	<b>Outlook: General MANOVA with Missing Data</b>	<b>28</b>

<b>7</b>	<b>Conclusions and Outlook</b>	<b>31</b>
	Bibliography	33
<b>II</b>	<b>Publications</b>	<b>38</b>

# List of Contributed Articles

This thesis is based on the following articles, referred to by their Roman numbers throughout the text.

- (I) Amro, L., and Pauly, M. (2017). Permuting incomplete paired data: a novel exact and asymptotic correct randomization test. *Journal of Statistical Computation and Simulation*, 87(6), 1148-1159, DOI: 10.1080/00949655.2016.1249871. *The reuse of this article in the thesis is granted by the copyright transfer agreement with the publisher.*

Contribution of the author:

The author of this thesis prepared and structured the manuscript with input from Prof. Pauly. Furthermore, she also implemented the simulation studies and conducted the analysis of the data example under Prof. Pauly's supervision.

- (II) Amro, L., Konietzschke, F., and Pauly, M. (2019). Multiplication-Combination Tests for Incomplete Paired Data. *Statistics in Medicine*, 38(17), 3243-3255, DOI: 10.1002/sim.8178. *The reuse of this article in the thesis is granted by the copyright transfer agreement with the publisher.*

Contribution of the author:

The author of this thesis had a leading role in drafting and structuring the paper with input from all coauthors. She also implemented the extensive simulation studies and conducted the analysis of the data example.

- (III) Amro, L., Pauly, M., and Ramosaj, B. (2021). Asymptotic-based bootstrap approach for matched pairs with missingness in a single-arm. *Biometrical Journal*, 63(7), 1389-1405, DOI: 10.1002/bimj.202000051. *The reuse of this article in the thesis is granted under the terms of the Creative Commons Attribution 4.0 International License.*

Contribution of the author:

The author of this thesis prepared and structured the paper mainly on her own. She mainly conducted the mathematical proofs and implemented the extensive simulation studies. Furthermore, she chose and analyzed the data examples independently.



- (IV) Amro, L., Konietschke, F., and Pauly, M. (2021). Incompletely observed nonparametric factorial designs with repeated measurements: A wild bootstrap approach. *Submitted to Journal of Multivariate Analysis*. Preprint: arXiv:2102.02871.

Contribution of the author:

The author of this thesis prepared and structured the manuscript mainly on her own. She conducted the mathematical proofs as well as the extensive simulation studies with helpful comments by the coauthors. Furthermore, she chose and analyzed the data examples independently.

## Further Published Papers:

- Ramosaj, B., Amro, L., and Pauly, M. (2020). A cautionary tale on using imputation methods for inference in matched-pairs design. *Bioinformatics*, *36(10)*, 3099-3106, DOI: [10.1093/bioinformatics/btaa082](https://doi.org/10.1093/bioinformatics/btaa082).
- Amro, L., and Samuh, M. H. (2017). More powerful permutation test based on multistage ranked set sampling. *Communications in Statistics-Simulation and Computation*, *46(7)*, 5271-5284, DOI: [10.1080/03610918.2016.1152364](https://doi.org/10.1080/03610918.2016.1152364).

# Abbreviations

ATS	ANOVA-type Statistic
MANOVA	Multivariate Analysis of Variance
MAR	Missing at Random
MATS	Modified ANOVA-type Statistic
MCAR	Missing Completely at Random
MNAR	Missing not at Random
WTS	Wald-type Statistic
$\mathbb{1}\{.\}$	Indicator function
$\mathbf{A}^\top$	Transpose of a matrix or (column) vector $\mathbf{A}$
$[\cdot]^+$	Moore-Penrose inverse of a matrix
$\xrightarrow{P}$	Convergence in probability
$tr(\mathbf{A})$	Trace of a matrix $\mathbf{A}$

**Part I**  
**Introduction**



# 1 Motivation

Repeated measures designs are frequently employed in a wide variety of fields, including medicine, psychology, biology, ecology, agriculture, and industry. This design refers to a study where the same subject is observed repeatedly under different experiment conditions or multiple time points. The subjects may be patients, cells, plants, animals, products, etc. Examples of repeated measures studies include the response of physiological and molecular traits of plants to environmental factors, progression of disease between treatment groups over time, annual revenue of companies across years, preoperative anxiety in women before and after surgery, and pollutant level change over time. Repeated measures design can be built up in an abundance of ways, from a simple set up of matched pairs design in which the same subject is observed twice to a more complex framework of mixed models or growth curves.

A remarkable merit of repeated measures design is that conducting several measurements on each subject eliminates within-subject variation, increases precision, and reduces error, thereby potentially increasing the study's power, for more details see Davis (2002). In spite of its desirable properties, repeated measure designs have a few critical challenges that must be addressed appropriately. First, since we have been measuring the same subject over and over again, the observations are correlated, and the assumption of the observation's independence is no longer tenable. Hence, data dependencies must be dealt with properly. Second, when it comes to long-term follow-up, obtaining completely observed data could be a fantasy. The latter was confirmed by Bell et al. (2014), who conducted a review of randomized clinical trials published between July and December 2013 in four top medical journals and discovered that 95% of them reported some missing outcome data. Consequently, complicated statistical models emerge, making the development of suitable statistical procedures challenging and limiting the number of available adequate techniques.

We will begin with the simplest repeated measures design, the matched pairs design. When there are missing data, the outcome is a mixture of paired components (complete cases) and unpaired observations (incomplete cases). Potential approaches that can handle mean-based inference for such data design are proposed in Bhoj (1978), Ekbohm (1976), Kim et al. (2005), Lin and Stivers (1974), Looney and Jones (2003), and Samawi and Vogel (2014). However, they all have the drawback of being developed under particular model assumptions such as bivariate normality, symmetry, or homoscedasticity that are difficult to verify in practice. If the assumptions are violated, inaccurate decisions due to potentially inflated or conservative type-I error

rates may occur (Amro and Pauly, 2017; Fong et al., 2018; Konietzschke et al., 2012; Samawi and Vogel, 2014; Xu and Harrar, 2012). Thus, we are interested in developing statistical methods for the problem of matched pairs with missing values that are robust against deviations from parametric assumptions and lead to valid inferences in case of heteroscedasticity or skewed distributions. Besides, we solve the following shortcomings of existing incomplete matched pairs approaches:

- Missing values do not necessarily occur on both arms but may occur on a single arm, and most available approaches are inapplicable under this unique missing pattern (Qi et al., 2019).
- Most of the available methods test for differences among the means. If ordinal or ordered categorical data are present, mean-based approaches are not applicable and show their limitation (Fong et al., 2018).

For this, we are addressing this incomplete paired data problem under both semi-parametric and nonparametric setups. Under a semiparametric model, we propose several statistical approaches that can handle metric data, where null hypotheses are formulated in terms of means, with missingness in both arms ((I) and (II)) and with missingness in a single arm (III). Furthermore, we propose several approaches under the nonparametric model, where we base inference on rank-statistics of the relative effects (II). Thus, our proposed procedures work for data with nonnormal or normal distributions, heteroscedastic and homoscedastic variances, ordinal, categorical, and continuous data, and unbalanced and balanced sample sizes.

Moreover, this dissertation also discusses a more complicated design that investigates the main and interaction effects between several factors; repeated measurements. Parametric mean-based procedures such as multivariate analysis of variance (MANOVA) or linear mixed models may be applied for analyzing data from such designs. However, owing to the restrictive distributional assumptions that these procedures require, such as continuity, multivariate normality, or special dependencies, their practical application is not always plausible (Arnau et al., 2012; Konietzschke et al., 2015). In addition, classical MANOVA approaches are not applicable for ordinal or ordered categorical data. In such cases, nonparametric rank-based approaches provide an attractive option due to their numerous advantages, which can be summarized as follows:

- (1) They are not based on stringent parametric assumptions about the underlying distribution such as normality or homoscedasticity.
- (2) They are applicable for continuous as well as non-continuous data, such as discrete or ordered categorical data.
- (3) They are robust to outliers.
- (4) They are invariant under monotonic transformations of the data.

- (5) Nonparametric methods often provide more accurate test decisions than parametric methods in case of nonnormal distribution, arbitrary covariance structures, or unbalanced experimental designs.

For more details, please refer to Breitung and Gourieroux (1997), Brunner et al. (2018), and Hettmansperger and McKean (2010). Several nonparametric techniques are available in the literature for handling factorial designs with repeated measurements data (Akritas, 2011; Akritas and Arnold, 1994; Akritas and Brunner, 1997; Brunner et al., 2017; Brunner and Puri, 2001; Friedrich et al., 2017a; Munzel and Brunner, 2000). However, all the aforementioned approaches are only applicable for fully observed data. Other nonparametric tests that can handle incomplete data are proposed by Brunner et al. (1999) and Domhof et al. (2002). But, their suggested Wald-type test needs large sample sizes to obtain accurate test decisions and their ANOVA-type test is in general not asymptotically correct and does not control type-I error accurately under small sample sizes. Thus, we developed novel nonparametric approaches that have all of the aforementioned properties (1-5), as well as the following:

- (6) They can be used for repeated measures data with missing values.

In the sequel, we provide motivating examples that demonstrate the need for adequate test procedures to deal with incompletely observed data. Each of the following examples represents a particular case that can be handled by one (or two) of our suggested methods.

## 1. Coyote DNA study

We consider a study conducted by Riordan (2012) and aimed at comparing two techniques for extracting DNA from coyote blood samples. One technique was the QIAGEN DNeasy Blood and Tissue Kit, while the other was the chloroform isoamyl alcohol method. The study includes 30 different coyotes. Due to time and cost constraints, the researcher selected 6 coyotes at random and tested their DNA using both techniques, 8 with the kit and 16 with chloroform. For more details about this study, please refer to Einsporn and Habtzghi (2013). In order to determine if the extraction methods produce a significantly different mean concentration of DNA, a method for testing mean based hypothesis is required. This method should be capable of handling small sample sizes and data with missingness in both arms. Thus, the weighted combination method proposed in (I) or the multiplication combination approach for testing mean hypotheses (II) is recommended.



## 2. Breast Cancer Study

The breast cancer study has been conducted by the Cancer Genome Atlas (TCGA) project to better diagnose, treat, and prevent breast cancer (Clark et al., 2013; Koboldt et al., 2012). The study involves 1093 breast cancer patients, and we were interested in the patients with pathological stage I. This subgroup consists of 90 patients. Sixteen provide both normal and tumor tissue, and 4 with only tumors. We study the following eight genes: TP53, ABCC1, HRAS, GSTM1, ERBB2, CD8A, C1D, and GBP3. We want to see whether early stage I breast cancer patients' mean genetic expressions of each gene significantly differ between normal and tumor tissues. We have here the situation of matched pairs with missingness in a single arm. We analyze this data example in detail in section 5 in (III).

## 3. Migraine study

We consider the clinical migraine study conducted by Kostecki-Dillon et al. (2018) which investigates four sessions of a nondrug headache treatment program. An ordinal scale ranging from 0 to 20 was used to assess the headache severity level over the treatment sessions. A total of 135 migraine patients took part in this study. However, the data contains a large number of missing observations. We only consider the patients' first and third session clinical records to fit in the matched pairs design. From the 132 patients involved in Session 1 and Session 3, only  $n_c = 82$  patients were measured twice,  $n_1 = 44$  patients were only seen in Session 1 and  $n_2 = 6$  patients were only assessed in Session 3. Since the observed data are ordinal grading scores and means do not offer an appropriate measure for score data, we need a rank-based matched pairs approach to investigate the effects of attending the sessions. Additionally, the considered approach must be capable of handling paired data with missingness in both arms. We analyze this data example in detail in (II).

## 4. Skin disorder trial

We consider the skin disorder trial published by Davis (2002) to investigate a skin condition's severe rate over time and compare two therapy treatments, drugs, and placebo. In this study, 88 patients received the drug active treatment, and 84 patients were in the placebo group. An ordinal response scale was used to evaluate the degree of improvement over three follow-up visits (1 = rapidly improving, 2 = slowly improving, 3 = stable, 4 = slowly worsening, 5 = rapidly worsening). This is a factorial design with two factors "treatment" and "time". This dataset was not complete, and approximately 30% of observations were missing. Due to the ordinal score data, nonparametric rank-based methods are, in this case, applicable. Further, the method should be able to handle data with a moderate number of missing values. Thus, we analyze the data using the factorial designs method proposed in (IV).

This dissertation is organized as follows: A brief overview of the consequences of missing data and the missing data mechanisms is presented in Chapter 2. Chapter 3 and 4 outline our different considered statistical models and resampling procedures, respectively. Chapter 5 provides summaries of the four articles this dissertation is based upon. While, Chapter 6 describes the fifth scenario; a semiparametric setting for multivariate data with missing values which is still under development. Lastly, Chapter 7 contains discussions of the results, conclusions, and an outlook to some future researches. The four articles and their supplementary materials are included in Part II of the dissertation.

## 2 Missing Data

In order to conduct statistical inference for decision making, a sufficient amount of data must exist that can be used to retrieve information. However, the collected data might be incomplete in many situations. Many factors can lead to data with missing values, including sample subjects' refusal to provide data for certain variables, experimental equipment failures, missed study visits, or data entry mistakes. Missing data may have a substantial impact on the statistical inferences drawn from the data (Allison, 2001; Little and Rubin, 2019). It is intensified in repeated measurements designs since a single missed observation might impact several subsequent observations on the same research subject. The optimal solution is to repeat the experiment in order to receive the complete data set. However, this is not always achievable, mainly when measurements must be taken at certain time intervals, there are not enough additional experimental subjects, or rerunning the experiment would be too expensive. Thus, this is an impractical solution, other approaches for resolving this issue are needed.

The most simple approach for handling data that has missing values is the complete case analysis. This approach restricts the data analysis on the fully observed subjects, i.e., subjects without any missing values. For example, if one or more observations for subject  $i$  are missing, then all observations for subject  $i$  are deleted from the data set. Convenience and straightforwardness are the main advantages of this method; yet, ignoring some available information might result in biased estimates, a loss of statistical power, and perhaps inaccurate decisions, except in a few cases (Allison, 2001; Enders, 2010; Little and Rubin, 2019).

An alternative approach to deleting subjects with missing values is imputation, in which no observable data is omitted, and missing values are replaced with suitable substitutes estimated from the existing data. Then, data analysis can be performed the same as if the dataset were complete. Even though the imputation approach appears easy to implement, it requires a careful selection of the appropriate imputation model based upon the challenges of each studied dataset (Ramosaj et al., 2020). It is worth noting that performing a different analysis on the same dataset often provides different results based on the chosen imputation model, implying that the analysis decision is subjective. According to Guo and Yuan (2017), imputation approaches are not suggested for small sample sizes since they can result in inaccurate test decisions, such as inflated type-I error or low power behavior. Thus, it is vital to develop statistical techniques capable of analyzing incomplete data without ignoring

or substituting for any values. Therefore, we are developing in this dissertation several statistical methods that analyze all observed data without excluding any subject or imputing any missing values.

Rubin (1976) identified three missing mechanisms for the data based on the relationship between the missing values and observed values. Let  $\mathbf{Y}$  denote a data set which can be decomposed into observed and unobserved portions  $\mathbf{Y} = (\mathbf{Y}_{obs}, \mathbf{Y}_{mis})$ . Let  $\mathbf{R}$  be a binary matrix whose components indicate whether  $\mathbf{Y}$  is observed or missing. The three missing data mechanisms are:

**1. Missing completely at random (MCAR):**

The probability of an observation being missing does not depend on the values of any observed or unobserved data, i.e.,  $P(\mathbf{R}|\mathbf{Y}_{obs}, \mathbf{Y}_{mis}) = P(\mathbf{R})$ . This implies that the conditional and marginal distributions can always be accurately estimated from the observed data.

**2. Missing at random (MAR):**

The probability of the missingness can depend on the observed data but not on the unobserved data, i.e.  $P(\mathbf{R}|\mathbf{Y}_{obs}, \mathbf{Y}_{mis}) = P(\mathbf{R}|\mathbf{Y}_{obs})$ . Thus, the missing data is due to an external effect, not the variable itself. Note that MCAR is a special case of MAR.

**3. Missing not at random (MNAR):**

The probability of the missingness can depend on the unobserved data, i.e.  $P(\mathbf{R}|\mathbf{Y}_{obs}, \mathbf{Y}_{mis}) \neq P(\mathbf{R}|\mathbf{Y}_{obs})$ . MNAR is also known as the non-ignorable case (Little and Rubin, 2019) since the missing observation is dependent on the outcome result.

For additional information on the various missing mechanisms, we refer to Little and Rubin (2019).

Throughout this dissertation, we assume MCAR when constructing the test statistics and developing their related theories. However, in the simulation studies, we investigate the effects of some MAR and MNAR scenarios on the test statistics performance under small sample sizes.

# 3 Statistical Models

## 3.1 Semiparametric Models

We consider a general matched pairs design given by independent and identically distributed (i.i.d.) random vectors  $\mathbf{X}_j = [X_{1j}, X_{2j}]^\top, j = 1, \dots, n$ . with mean vector  $\mathbb{E}[\mathbf{X}_1] = \boldsymbol{\mu} = [\mu_1, \mu_2]^\top \in \mathbb{R}^2$  and an arbitrary positive definite covariance matrix  $Cov(\mathbf{X}_1) = \boldsymbol{\Sigma}$ .

### 3.1.1 Matched Pairs with Missing Values in Both Arms

We extend the general matched pairs design in that we allow for missing values in both components of the pair. To accommodate missing values, let  $\mathbf{R}_j = [R_{1j}, R_{2j}]^\top$  indicate whether  $X_{ij}$  is observed ( $R_{ij} = 1$ ) or missing ( $R_{ij} = 0$ ) for  $i = 1, 2, j = 1, \dots, n$ . Define the composition  $*$  by  $a * 1 = a$  and  $a * 0 = \text{---}$ , for all  $a \in \mathbb{R}$ , then we only observe  $\mathbf{X}^{(o)} := \{\mathbf{X}_j * \mathbf{R}_j\}_{j=1}^n$ , and a " --- " entry is interpreted as missing. Hence our framework has the following form:

$$\underbrace{\begin{bmatrix} X_{11}^{(c)} \\ X_{21}^{(c)} \end{bmatrix}, \dots, \begin{bmatrix} X_{1n_c}^{(c)} \\ X_{2n_c}^{(c)} \end{bmatrix}}_{\mathbf{X}^{(c)}} \cdot \underbrace{\begin{bmatrix} X_{11}^{(i)} \\ \text{---} \end{bmatrix}, \dots, \begin{bmatrix} X_{1n_1}^{(i)} \\ \text{---} \end{bmatrix}}_{\mathbf{X}^{(i)}} \cdot \underbrace{\begin{bmatrix} \text{---} \\ X_{21}^{(i)} \end{bmatrix}, \dots, \begin{bmatrix} \text{---} \\ X_{2n_2}^{(i)} \end{bmatrix}}_{\mathbf{X}^{(i)}}. \quad (3.1)$$

We assume that the first components  $X_{1j}^{(c)}, X_{1k}^{(i)}$  are i.i.d. with mean  $\mu_1$  and variance  $\sigma_1^2 \in (0, \infty)$  and the second components  $X_{2j}^{(c)}, X_{2l}^{(i)}$  are i.i.d. with mean  $\mu_2$  and variance  $\sigma_2^2 \in (0, \infty)$  for  $j = 1, \dots, n_c, k = 1, \dots, n_1, l = 1, \dots, n_2$ . And, the complete pairs  $[X_{1j}^{(c)}, X_{2j}^{(c)}]^\top$  are i.i.d. with mean vector  $\boldsymbol{\mu} = [\mu_1, \mu_2]^\top$  and an arbitrary positive definite covariance matrix  $\boldsymbol{\Sigma}$ . So, the covariance matrix allows for unequal variances (heteroscedasticity), and we do not assume any special covariance structure or any particular underlying distribution of the data. In article (I) we even discuss how to loosen these model assumptions to the case where the incomplete observations are only assumed to be independent (i.e. MAR settings).

In this set-up, we use all the available data to test the null hypothesis  $H_0 : \{\mu_1 = \mu_2\}$  against the one-sided alternative  $\{\mu_1 > \mu_2\}$  or the two-sided alternative  $\{\mu_1 \neq \mu_2\}$ .

We propose two novel tests under this semiparametric model: a weighted permutation test (WPT) (I) and multiplication combination test (MCT) (II). Based on our theoretical results and simulation study, our novel WPT and MCT are asymptotic valid and even finitely exact if specific invariance properties are met. They also exhibit favorable small sample properties and robust error control against heteroscedasticity or skewed distributions in most examined scenarios. This all makes our WPT and MCT applicable to our general model 3.1. However, the WPT often showed the better power behaviour and is recommended in general.

### 3.1.2 Matched Pairs with Missing Values in a Single Arm

This section addresses another setting in which just one of the matched pairs data components is missing. This situation might occur in cancer research in which some participants contribute both tumor and normal tissues while others provide just tumor tissues owing to, for example, normal tissue scarcity.

Thus, we extend the model at the onset of Section 3.1 to allow for missingness in one arm (say the second). To incorporate missing values, denote with  $\zeta_{2j} \in \{0, 1\}$ ,  $j = 1, \dots, n$  the vector whose  $j$ -th component indicates whether  $X_{2j}$  is observed ( $\zeta_{2j} = 1$ ) or missing ( $\zeta_{2j} = 0$ ) for  $j = 1, \dots, n$ . Define the composition  $*$  by  $a * 1 = a$  and  $a * 0 = \text{---}$ , for all  $a \in \mathbb{R}$ , then we observe  $\mathbf{X}^{(o)} := \{\mathbf{X}_j * \zeta_j\}_{j=1}^n$  where  $\zeta_j = [1, \zeta_{2j}]^\top \in \mathbb{R}^2$ ,  $j = 1, \dots, n$ , and a "---" entry is interpreted as missing. Hence our framework has the following form:

$$\underbrace{\begin{bmatrix} X_{11}^{(c)} \\ X_{21}^{(c)} \end{bmatrix}, \dots, \begin{bmatrix} X_{1n_c}^{(c)} \\ X_{2n_c}^{(c)} \end{bmatrix}}_{\mathbf{X}^{(c)}}, \underbrace{\begin{bmatrix} X_{11}^{(i)} \\ \text{---} \end{bmatrix}, \dots, \begin{bmatrix} X_{1n_1}^{(i)} \\ \text{---} \end{bmatrix}}_{\mathbf{X}^{(i)}}. \quad (3.2)$$

We assume that the first components  $X_{1j}^{(c)}, X_{1k}^{(i)}$  are i.i.d. with mean  $\mu_1$  and variance  $\sigma_1^2 \in (0, \infty)$  for  $j = 1, \dots, n_c, k = 1, \dots, n_1$ . And, the complete pairs  $[X_{1j}^{(c)}, X_{2j}^{(c)}]^\top$  are i.i.d. with mean vector  $\boldsymbol{\mu} = [\mu_1, \mu_2]^\top$  and some unstructured covariance matrix  $\boldsymbol{\Sigma} > 0$  that allows for heteroscedastic variances. Let  $N = 2n_c + n_1$  denote the total number of observations and  $n = n_c + n_1$  the total number of subjects.

In this setting, we would like to test the null hypothesis  $H_0 : \{\mu_1 = \mu_2\}$  against the alternative hypothesis  $H_1 : \{\mu_1 \neq \mu_2\}$ . Our statistical model has the advantage of incorporating all the available data. It also drops the common parametric assumption such as homoscedasticity and normality. To derive the asymptotic results, we only assume that the following convergences

- $\frac{n_c}{n_c + n_1} \rightarrow \kappa_1 \in (0, 1)$ ,

- $\frac{n_1}{n_c+n_1} \rightarrow \kappa_2 = (1 - \kappa_1) \in (0, 1)$ ,

hold as  $\min\{n_c, n_1\} \rightarrow \infty$ . We propose three asymptotic model based bootstrap tests based upon the quadratic form test statistics: WTS, ATS, and MATS (III). Our novel tests are asymptotically correct and robust under heteroscedasticity and skewed distributions in most considered scenarios, which make them applicable to our general model 3.2.

## 3.2 Nonparametric Models

### 3.2.1 Matched Pairs with Missing Values in Both Arms

In order to deal with situations involving ordinal or ordered categorical data where means are inadequate measures, we consider a purely nonparametric model, more general than in Section 3.1. Hence, we assume Model (3.1) with arbitrary unknown marginal distribution functions  $F_i$  for component  $i = 1, 2$ . Only the trivial assumption that  $F_i$  is not a one point distribution is required. As with the previous models, this model does not discard any available information from the data; rather, it incorporates all available data.

This model does not contain any parameter to describe differences between the marginal distributions, thus we consider the WMW-effect

$$p = \int F_1 dF_2 = P(X_{11}^{(c)} < X_{22}^{(c)}) + 1/2P(X_{11}^{(c)} = X_{22}^{(c)}), \quad (3.3)$$

also known as the (nonparametric) treatment effect or relative marginal treatment effect, (Brunner and Munzel, 2000; Brunner and Puri, 1996; Fligner and Policello, 1981; Munzel, 1999). The interpretation of the relative effect in a matched pairs design is as follows: If  $p > 1/2$ , the observations with distribution  $F_2$  tend to be larger than those with distribution  $F_1$ , and vice versa if  $p < 1/2$ . Moreover,  $p = 1/2$  applies to the case of no treatment effect. Thus, we consider null hypothesis formulated in terms of the relative effect measure  $p$  as  $H_0^p : \{p = 1/2\}$ . This hypothesis is less restrictive than the hypothesis  $H_0^F : \{F_1 = F_2\}$  since  $F_1 = F_2$  implies  $p = 1/2$  but not vice versa (Brunner et al., 2017).

In addition to this nonparametric Behrens-Fisher problem, we also consider two more testing problems under this nonparametric model

- Wilcoxon rank sum procedures: We assume a shift model with  $F_1(x) = F_2(x - \delta)$  for some  $\delta \geq 0$ . The null hypothesis is formulated as  $H_0^\delta : \{\delta = 0\}$  which may be tested against the one-sided alternative  $\{\delta > 0\}$ .
- Test procedures for  $H_0^F : \{F_1 = F_2\}$ .

Three distinct multiplication combination tests are proposed to address each of the aforementioned testing problems in this nonparametric model (II). Theoretically and numerically, we demonstrate that our proposed approaches are asymptotically valid and even finitely exact when specific invariance properties are met.

### 3.2.2 Repeated Measures with Missing Values

This section discusses a more broad design than above that takes into consideration both the main and interactions effects between several variables. We consider a repeated measures model consisting of  $a$  independent and possibly imbalanced treatment groups with  $d$  different time points given by independent random vectors

$$\mathbf{X}_{ik} = [X_{i1k}, \dots, X_{idk}]^\top, \quad i = 1, \dots, a; \quad k = 1, \dots, n_i, \quad (3.4)$$

with marginal distributions  $X_{ijk} \sim F_{ij}(x)$ ,  $i = 1, \dots, a; \quad j = 1, \dots, d$ .

Ordinal observations in a multi-group repeated measures design are not uncommon. Hence, to allow for continuous, discrete, and even dichotomous data, we employ the normalized version of the distribution function

$$F_{ij}(x) = \frac{1}{2}[F_{ij}^+(x) + F_{ij}^-(x)], \quad (3.5)$$

where  $i = 1, \dots, a, j = 1, \dots, d, k = 1, \dots, n_i$ . Here,  $F_{ij}^+(x) = P(X_{ijk} \leq x)$  and  $F_{ij}^-(x) = P(X_{ijk} < x)$  are the right- and left-continuous versions of the distribution function  $F_{ij}(x)$ , respectively (Brunner et al., 2018). Note that  $F_{ij}$  may be arbitrary distributions, with the exception of the case of one point distribution.

We extend the classical nonparametric repeated measures design in which we allow for missing values. To include missingness, let  $\lambda_{ijk}$ , a missing indicator, as

$$\lambda_{ijk} = \begin{cases} 1, & \text{if } X_{ijk} \text{ is observed} \\ 0, & \text{if } X_{ijk} \text{ is non-observed} \end{cases} \quad i = 1, \dots, a; \quad j = 1, \dots, d; \quad k = 1, \dots, n_i. \quad (3.6)$$

Let  $n = \sum_{i=1}^a n_i$  denote the total number of subjects and  $N = \sum_{i=1}^a \sum_{j=1}^d \sum_{k=1}^{n_i} \lambda_{ijk}$  the total number of observations.

Our general model does not include any parameter that describes the differences between the marginal distributions. Hence, we consider the relative treatment effects

$$p_{ij} = \int H(x) dF_{ij}(x), \quad (3.7)$$



where  $H(x) = N^{-1} \sum_{i=1}^a \sum_{j=1}^d \sum_{k=1}^{n_i} \lambda_{ijk} F_{ij}(x)$  denotes the weighted mean of all distribution functions (Brunner et al., 1999; Domhof et al., 2002).

In this nonparametric setup, the null hypotheses are formulated by  $H_0 : \{\mathbf{C}\mathbf{F} = \mathbf{0}\}$ , where  $\mathbf{F} = [F_{11}, \dots, F_{ad}]^\top$  denotes the vector of the distribution functions  $F_{ij}$ ,  $i = 1, \dots, a$ ,  $j = 1, \dots, d$ . and  $\mathbf{C}$  denote a contrast matrix, i.e.,  $\mathbf{C}\mathbf{1} = \mathbf{0}$  where  $\mathbf{1} = [1, \dots, 1]^\top$  and  $\mathbf{0} = [0, \dots, 0]^\top$ . Let  $\mathbf{P}_d = \mathbf{I}_d - \frac{1}{d}\mathbf{J}_d$  be the d-dimensional centering matrix, where  $\mathbf{I}_d$  is the d-dimensional identity matrix and  $\mathbf{J}_d$  is the  $d \times d$  matrix of  $\mathbf{1}$ 's i.e.  $\mathbf{J}_d = \mathbf{1}_d \mathbf{1}_d^\top$ , where  $\mathbf{1}_d = [1, \dots, 1]_{d \times 1}^\top$  denotes the d-dimensional column vector. This framework is applicable to various factorial repeated measures designs. It also covers many null hypotheses of interest such as

$$\begin{aligned} H_0^G &: \{(\mathbf{P}_a \otimes \frac{1}{d}\mathbf{1}_d^\top)\mathbf{F} = \mathbf{0}\} && \text{(no treatment group effect),} \\ H_0^T &: \{(\frac{1}{a}\mathbf{1}_a^\top \otimes \mathbf{P}_d)\mathbf{F} = \mathbf{0}\} && \text{(no time effect),} \\ H_0^{GT} &: \{(\mathbf{P}_a \otimes \mathbf{P}_d)\mathbf{F} = \mathbf{0}\} && \text{(no interaction effect between treatment and time),} \end{aligned}$$

where we denote by  $\mathbf{A} \otimes \mathbf{B}$  the Kronecker product of the matrices  $\mathbf{A}$  and  $\mathbf{B}$ .

In order to derive asymptotic results, we assume the following sample size and missing values assumption:

$$\frac{\lambda_{ij.}}{n} \rightarrow \kappa_i \in (0, 1) \quad i = 1, \dots, a; \quad j = 1, \dots, d, \quad (3.8)$$

as  $\min\{\lambda_{ij.}\} \rightarrow \infty$ .

We propose in (IV) three asymptotically correct testing procedures using a wild bootstrap approach based upon WTS and ATS (Brunner et al., 1999; Domhof et al., 2002) as well as MATS (Friedrich and Pauly, 2018). Our simulations show their applicability under this nonparametric model for repeated measures with incomplete data.

## 4 Resampling Methods

Resampling methods are frequently employed as a more robust alternative to standard statistical inference techniques (Peterson et al., 2010). In comparison with standard parametric statistical methods, resampling tests have the following two advantages. First, they do not require strict parametric assumptions like variance homogeneity or normally distributed error terms. Second, they are often more robust and flexible than classical parametric methods. Additionally, resampling testing methods are appropriate and could be unavoidable when, e.g., classical parametric tests assumptions are violated, very complicated formulas are required for obtaining critical values, or no adequate parametric methods are available to accomplish our goals. For example, conducting an analysis using a parametric test when some of its assumptions are violated may result in misleading decisions or less powerful tests than the resampling tests counterparts.

The main idea of the resampling methods is to draw statistical decisions through the artificial resampling of the data. Thus, it is based upon repeatedly and randomly shuffling (or arranging) the data using a specific resampling approach and computing the test statistic at each resample process and for the original data. Then, this resampling probability mechanism is used to estimate the unknown theoretical distribution of the statistic of interest. The resampling test is (at least) asymptotically valid as long as the conditional resampling distribution asymptotically coincides with the corresponding distribution of the test statistic under the null.

The algorithm for the computation of the p-value based on a resampling distribution is as follows:

1. For the given data, calculate the observed test statistic, say  $T$ .
2. Resample the data according to the considered resampling procedure.
3. Calculate the value of the test statistic for the resampled sample  $T^*$ .
4. Repeat Steps 2 and 3 independently  $B = 999$  times and collect the observed test statistic values in  $T_b^*, b = 1, \dots, B$ .
5. Finally, estimate the resampling p-value as  $\text{p-value} = \frac{\sum_{b=1}^B \mathbb{1}\{T_b^* \geq T\}}{B}$ .

Several resampling procedures are available in the literature, such as the bootstrap, jackknife, and permutation test (Good, 2006; Shao and Tu, 2012). The following sections explain the three resampling techniques used in this dissertation: permutation, parametric bootstrap, and wild bootstrap.

## 4.1 Permutation Test

The permutation test (sometimes called a randomization test or an exact test (Good, 2005)) was first introduced by Fisher (1935) and later extended by Pitman (1937). The objective of permutation tests is to generate the sampling distribution of a test statistic from the values obtained by computing the test statistic under the null hypothesis for all possible permutations of the data.

In the last years, permutation tests have been widely applied to matched pairs designs, factorial designs, and multivariate designs, among other designs (Friedrich et al., 2017a; Janssen, 1999; Pauly et al., 2015; Pesarin and Salmaso, 2012; Salmaso, 2015). Here, we apply the permutation method both in the semiparametric matched pairs setting (I) and (II) and in the nonparametric matched pairs setting (II).

For the semiparametric matched pairs setting, several researchers have even used permutation tests (Maritz, 1995; Yu et al., 2012). However, to get a (at least asymptotically) valid level test, these approaches require specific distributional assumptions such as 0-symmetry, equal variances, or sample sizes. Thus, we propose two novel permutation tests that are asymptotically valid and even finitely exact under certain invariance properties while also being (asymptotically) robust to deviations like heteroscedasticity or skewed distributions (WPT (I) and MCT (II)). They are based upon combining independent results from paired and unpaired studentized permutation tests. For the completely observed case, we consider a studentized permutation test in the paired  $t$ -test, where each pair's components are permuted at random, as recommended by Janssen (1999) and Konietschke and Pauly (2014). For the incompletely observed case, we consider a studentized permutation test in the Welch-type statistic that is based on randomly permuting the pooled sample, as recommended by Janssen (1997) and Janssen (2005) as well as Janssen and Pauls (2003).

In the nonparametric setup, we propose a multiplication combination approach for testing the general Behrens-Fisher problem for incompletely observed paired data. Also, we exemplify the adaptability of our approach for two other testing problems: Wilcoxon rank sum procedures and testing  $H_0^F : \{F_1 = F_2\}$ . The developed tests are available in our paper (II).

## 4.2 Bootstrap

Efron (1979) proposed the bootstrap technique as a computer-based resampling approach to estimate the standard error of a parameter estimate. These-day, the bootstrap is widely applied to a variety of statistical procedures. One of the main benefits of the bootstrap approach is summarized by Efron and Tibshirani (1994) as follows: "The bootstrap can answer questions which are too complicated for traditional statistical analysis." In the following sections, we provide an overview of the two bootstrap methods used in this thesis; parametric bootstrap and wild bootstrap.

### 4.2.1 Parametric Bootstrap

We use an asymptotic model-based bootstrap method to estimate critical values. This approach has, e.g., previously been used in the context of (M)ANOVA factorial designs (Friedrich and Pauly, 2018; Konietzschke et al., 2015). We generate a parametric bootstrap sample as follows:

$$\mathbf{X}_j^* = \begin{bmatrix} X_{1j}^* \\ X_{2j}^* \end{bmatrix} \stackrel{\text{i.i.d.}}{\sim} N(0, \hat{\mathbf{\Gamma}}), \quad j = 1, \dots, n. \quad (4.1)$$

Here,  $\hat{\mathbf{\Gamma}} = \begin{bmatrix} \hat{\sigma}_1^2 & \hat{\rho}\hat{\sigma}_1\hat{\sigma}_2 \\ \hat{\rho}\hat{\sigma}_1\hat{\sigma}_2 & \hat{\sigma}_2^2 \end{bmatrix}$  is the empirical covariance matrix estimator, i.e.  $\hat{\sigma}_i^2$  denotes the sample variance calculated on all observations in component  $i$  and  $\hat{\rho}$  is the sample correlation obtained from  $\mathbf{X}^{(c)}$ .

The main idea is to mimic the covariance structure of the data to get a more accurate finite sample approximation. For the semiparametric setting for matched pairs with incompleteness in a single arm, we propose three asymptotic model-based bootstrap tests that are not based on any parametric assumptions and use all observed information. The tests were based upon cleverly restructuring all available information in quadratic form test statistics; WTS, ATS, and MATS. We show that under both the null and alternative hypotheses, the conditional distribution of all three bootstrapped test statistics, given the data, weakly converges to the null distribution of the respective test statistic in probability ((III), Theorem 3.1 in the supplement). Thus, the parametric bootstrap tests are asymptotically correct resampling procedures, and they are robust under heteroscedasticity and skewed distributions.

### 4.2.2 Wild Bootstrap

The wild bootstrap has been extensively studied and applied in the literature, specifically, for MANOVA and repeated measurements (Friedrich et al., 2017b; Friedrich and Pauly, 2018; Konietzschke et al., 2015; Xu et al., 2013). We generate wild bootstrap samples as follows: First, let  $W_{ik}$  denote independent and identically distributed random weights with  $E(W_{ik}) = 0$  and  $Var(W_{ik}) = 1$ . There are several choices for these random weights (Davidson and Flachaire, 2008; Mammen, 1992). We employ Rademacher random variables, which are specified by  $P(W_{ik} = -1) = P(W_{ik} = 1) = 1/2$ . Then, a wild bootstrap sample is generated by multiplying the fixed data with Rademacher random weights. For the nonparametric factorial design setting, we apply the wild bootstrap and generate the bootstrapped sample as follows:

$$\mathbf{Z}_{ik}^* = W_{ik} \mathbf{Z}_{ik}, \quad i = 1, \dots, a; \quad k = 1, \dots, n_i, \quad (4.2)$$

i.e.  $\mathbf{Z}_{ik}^*$  is a symmetrization of the rank vector  $\mathbf{Z}_{ik}$  (defined in (IV)). Based on these bootstrap variables, we construct three asymptotically valid bootstrap tests based upon the Wald, ANOVA, and modified ANOVA type test statistics. According to theorem 4.2 in (IV), the wild bootstrap tests have asymptotic level  $\alpha$  under the null hypothesis and are consistent for any fixed alternative  $H_1 : \{\mathbf{CF} \neq \mathbf{0}\}$ , i.e., they have asymptotic power 1. Furthermore, it shows that they have the same local power under contiguous alternatives as their original tests (Jansen et al., 2003).

# 5 Summary of the Scientific Articles

## 5.1 Article I: Permuting Incomplete Paired Data: a novel exact and asymptotic correct randomization test.

In this article, we consider the semiparametric model for matched pairs data with missing values in both arms with possibly non-normal and/or heteroscedastic data stated in Section 3.1.1. In such settings, most existing procedures demonstrate their limitation due to their distributional assumptions such as normality, 0-symmetry, or homoscedasticity of the data. If these particular assumptions are not fulfilled, the corresponding tests' type-I error control may be inflated. The main idea of this article was to develop a novel randomization approach for partially paired data that does not depend on any parametric assumptions, uses all observed data, and is robust against deviations such as heteroscedasticity or skewed distributions, i.e., valid for our general model. In this approach, separate results from studentized permutation tests for the paired and unpaired cases are combined. This was accomplished by merging independent results about studentized permutation tests for the paired and unpaired cases. Hence, we develop a general weighted test statistic

$$T = T(\mathbf{X}^{(c)}, \mathbf{X}^{(i)}) = \sqrt{a}T_1(\mathbf{X}^{(c)}) + \sqrt{1-a}T_2(\mathbf{X}^{(i)}), \quad (5.1)$$

that combines both the paired t-test statistic  $T_1$  calculated for the complete observations  $\mathbf{X}^{(c)}$  and the Welch test statistic  $T_2$  computed for the incomplete observations  $\mathbf{X}^{(i)}$ . Here the weighting coefficient  $a \in [0, 1]$  is used to assess the impact of complete observations on incomplete observations. However, in cases of small sample sizes or skewed distributions, the estimate of test statistic  $T$  may be inaccurate, resulting in either liberal or conservative test decisions. Therefore, we propose a permutation version of the weighted test that is based upon a studentized permutation test in the paired t-test  $T_1$ , and a studentized permutation test in the Welch test  $T_2$  i.e.

$$T_p = T(\mathbf{X}_\tau^{(c)}, \mathbf{X}_\pi^{(i)}) = \sqrt{a}T_1(\mathbf{X}_\tau^{(c)}) + \sqrt{1-a}T_2(\mathbf{X}_\pi^{(i)}), \quad (5.2)$$

where  $\mathbf{X}_\tau^{(c)}$  is the permuted vector which is based on randomly permuting the

components of the complete sample  $\mathbf{X}^{(c)}$ , and  $\mathbf{X}_\pi^{(i)}$  is the permuted vector which is based upon randomly permuting all elements of the pooled incomplete sample  $\mathbf{X}^{(i)}$ .

We proved that the conditional distribution of the permutation version of the test statistic  $T_p$  always approximates the null distribution of  $T$  given the data, thus leading to an asymptotic level  $\alpha$  test, which is even finitely exact if certain invariance properties are fulfilled (see Theorem 3.1 in (I)). In order to investigate the finite sample sizes behavior of our weighted permutation test, we conducted an extensive simulation study in which we also compared our method to three other approaches. We have generated partially paired data under the MCAR framework using various symmetric and skewed bivariate distributions under a homoscedastic and a heteroscedastic covariance structure and different sample sizes. The simulation results revealed that our randomization test has favorable small sample properties that outperforms the alternative approaches and improves the small sample behavior of the asymptotic test based on the same weighted test statistic  $T$ . The simulation also indicated that our novel test provides an adequate level  $\alpha$  test for homoscedasticity as well as heteroscedasticity in almost all considered scenarios. Except for skewed exponential distribution with large correlations, the type-I error control is not adequate. However, all other considered approaches failed to control the type-I error rate in this scenario. Moreover, our proposed studentized permutation test has the largest power under all studied scenarios.

In addition, we consider a data example from a hospice clinical study. Therein, we aimed to compare the Karnofsky Performance Status Scale (KPS) results of the patients on the day before they died and on their last day in the life. There were a total of  $n_c = 9$  complete pairs and two unpaired samples of sizes of  $n_1 = 28$  and  $n_2 = 23$ . Hence, a large proportion of missing values and a small proportion of complete pairs in this scenario. The results indicated a significant difference between the mean KPS scores with respect to the patients' last two days in life.

Finally, we could confirm that our suggested permutation test is asymptotically correct, finitely exact under invariance, and possess favorable small sample properties. Also, our permutation test is robust to deviations from parametric assumptions while still allowing for adequate inference in the presence of heteroscedasticity or skewed distributions.

## 5.2 Article II: Multiplication-Combination Tests for Incomplete Paired Data.

In this work, we have focused on matched pairs with missing values in both arms while aiming to develop statistical methods for testing hypotheses formulated in terms of real-valued functionals. Most tests in the literature are based on parametric or semiparametric mixed models that involve difficult-to-verify assumptions like symmetry or bivariate normality. Besides, these methods are generally nonrobust to deviations, which might lead to incorrect decisions. The key objective was to provide a flexible approach that can be applied for parametric, semiparametric, and nonparametric models and can be used to test a variety of distinct hypotheses of interest. Therefore, we developed a novel multiplication-combination procedure to overcome the limitations of existing approaches and enhance their accuracy. The idea is to divide the observed data into completely observed pairs  $\mathbf{X}^{(c)}$  and incompletely observed components  $\mathbf{X}^{(i)}$  as in model 3.1. Then, assuming a MCAR mechanism and denoting with  $\varphi^{(c)} = \varphi^{(c)}(\mathbf{X}^{(c)})$  and  $\varphi^{(i)} = \varphi^{(i)}(\mathbf{X}^{(i)})$  adequate tests for the null hypothesis of interest that are computed upon  $\mathbf{X}^{(c)}$  and  $\mathbf{X}^{(i)}$  separately. Thus, by calculating each test at significance level  $\alpha^{1/2}$  individually, we can state a level  $\alpha \in (0, 1)$  multiplication-combination test (MCT) by

$$\varphi = \varphi^{(c)} \cdot \varphi^{(i)}. \quad (5.3)$$

The main advantage of this approach is that estimating  $(1 - \alpha^{1/2})$ -quantiles (of the underlying test statistics) is generally more accurate than estimating the common  $(1 - \alpha)$ -quantiles. Thus, an enhanced type-I error control is expected in the case of small to moderate sample sizes. We proved the validity of the MCT, resulting in an asymptotic level  $\alpha$ -test that is even finitely exact if certain invariance properties are met (see Theorem 1 in (II)). To demonstrate the adaptability of our multiplication approach, we proposed the following testing procedures:

1. A semiparametric procedure for mean comparisons: under the semiparametric model described in Section 3.1, we developed a procedure for testing  $H_0^\mu : \{\mu_1 = \mu_2\}$  (Section 2 in (II)).
2. Nonparametric rank-based test procedures: under the nonparametric model stated in Section 3.2, we proposed Wilcoxon rank sum procedures for testing  $H_0^\delta : \{\delta = 0\}$  for some  $\delta \geq 0$  with an assumed shift model with  $F_1(x) = F_2(x - \delta)$ , test procedures for  $H_0^F : \{F_1 = F_2\}$ , and test procedures for the nonparametric Behrens-Fisher problem  $H_0^p : \{p = 1/2\}$  (Section 3 in (II)).

In order to investigate the small sample behavior of our MCT, we have conducted an extensive simulation study. The complete pairs have been generated using symmetric as well as skewed distributions under homoscedastic and heteroscedastic settings. Missing values have been artificially inserted under the MCAR, MAR, and MNAR



mechanisms with various missing rates. Our simulation results indicated that our novel MCT approaches are more accurate than the alternative approaches in most considered scenarios, particularly for the nonparametric Behrens-Fisher problem.

In addition to the simulation study, we considered the clinical migraine study provided in Chapter 1. Therein, we aimed to compare the headache severity level of the patients in their first and third sessions. The clinical records of the patients were incomplete and based on ordinal scores. Thus, we analyzed the data using our proposed nonparametric rank-based test procedure. The results indicated a significant difference between the two sessions' headache severity levels.

Finally, we could conclude that the MCT can be used in various statistical models, including semiparametric and completely nonparametric models. A MCT based on permutation versions of the Munzel (1999) (complete case) and Brunner and Munzel (2000) (incomplete case) tests significantly improved the type-I error control of existing approaches for testing the Behrens-Fisher problem  $H_p : \{p = 1/2\}$ . Surprisingly, less restrictive missing mechanisms such as MAR and MNAR have little impact on the MCT type-I error control. However, the power behavior is strongly affected, which makes it more applicable under the MCAR framework.

## 5.3 Article III: Asymptotic Based Bootstrap Approach for Matched Pairs with Missingness in a Single-arm.

The problem of paired data with missing values is concerned not only about the occurrence of missing values but also about their location. Many approaches are available for testing hypotheses, but most are inapplicable when missing values happen in a single arm. Hence, these approaches cannot analyze data from The Cancer Genome Atlas (TCGA) project on pathological stage I breast cancer patients. This data set contains observations from 90 patients, 74 of them had entries in one component, although only 16 of them were complete; for more details, see Section 7.1 in (III). There is barely any work that applies to paired data with missingness in a single arm, needs no parametric assumptions, and leads to accurate decisions in the presence of heteroscedasticity or skewed distributions. An exception is the recent methods by Qi et al. (2019) who suggested the so-called nonparametric combination test (NCT) and nonparametric P-value pooling methods (NPM). However, our simulation studies revealed that the NCT and NPM might result in strongly inflated type-I error rate or significant power loss e.g., in the case of heteroscedasticity and/or skewed distributions.

Thus, the aim was to develop statistical tests that can handle single arm missing values in matched pairs while not losing (any) information, not based upon parametric assumptions like homoscedasticity and normality, and robust under heteroscedasticity and skewed distributions. Thus, we proposed three test statistics, WTS, ATS, and MATS, for our design. However, the WTS slowly converges to its limiting distribution under small sample sizes. In addition, the ATS and MATS are not distribution free under  $H_0$ , and critical values cannot be directly calculated. Therefore, we developed the parametric bootstrap WTS  $T_W^*$ , ATS  $T_A^*$ , and MATS  $T_M^*$  based on the mean differences vectors  $\mathbf{AZ}_n^*$  and empirical covariance matrices  $\hat{\Sigma}_n^*$  of the bootstrapped observations as

$$T_W^* = [\mathbf{AZ}_n^*]^\top [\mathbf{A}\hat{\Sigma}_n^*\mathbf{A}^\top]^{-1} [\mathbf{AZ}_n^*], \quad (5.4)$$

$$T_A^* = \frac{1}{tr(\mathbf{A}\hat{\Sigma}_n^*\mathbf{A}^\top)} [\mathbf{AZ}_n^*]^\top [\mathbf{AZ}_n^*], \quad (5.5)$$

$$T_M^* = [\mathbf{AZ}_n^*]^\top \hat{\mathbf{D}}_n^* [\mathbf{AZ}_n^*], \quad (5.6)$$

where  $\mathbf{A} = \begin{bmatrix} 1 & -1 & 0 \\ 0 & -1 & 1 \end{bmatrix}$ , and  $\hat{\mathbf{D}}_n^* = diag([\mathbf{A}\hat{\Sigma}_n^*\mathbf{A}^\top]_{ii}^+)$ . We proved that all three aforementioned bootstrapped test statistics  $T_W^*$ ,  $T_A^*$ , and  $T_M^*$  approximate the null distribution of the respective test statistic (Theorem 3.1 in the supplement of (III)). In addition, we have conducted an extensive simulation study to investigate the finite sample behavior of our suggested approaches for symmetric and skewed distributions

with homoscedastic and heteroscedastic covariance settings. We have considered several missing mechanisms and compared our approaches to three other alternative approaches (Little, 1976; Pesarin and Salmaso, 2010; Qi et al., 2019). Moreover, we developed a parametric bootstrap version of the test of Little (1976) to improve its small sample properties, which we investigated in our simulation study. Our findings revealed that the parametric bootstrap versions of WTS, ATS, MATS, and Little enhanced their asymptotic tests' small sample behavior. Our bootstrap tests performed very well under almost all scenarios, even under a large number of missing values, heteroscedastic covariance, or skewed data. Only the type-I error control is not adequate for the case of the exponential distribution, in particular, when heteroscedasticity, and MCAR, combined with a small number of complete pairs and a large number of incomplete observations. However, all alternative approaches failed to control the type-I error rate under this scenario. The bootstrap methods' type-I error control was unaffected by less stringent missing data mechanisms such as MAR. However, their power behavior was affected. This could be due to the related dependency structure in comparison to the MCAR.

In addition, we investigated three real-life examples from different fields and sources. We considered breast cancer study, anorexia study, and grapefruit study. First, the breast cancer study aimed to identify the genes that are significantly associated with breast cancer. We considered a data set containing  $n_c = 16$  complete pairs (normal and tumor), and  $n_1 = 74$  incomplete pairs (only tumor tissues). The anorexia study, on the other hand, aimed to compare the weights of 17 girls before and after anorexia treatment. Moreover, the grapefruit study was designed to investigate significant differences between shaded and exposed grapefruits. The anorexia and grapefruit datasets were complete. Thus, we have introduced 30% missing values on each data set. The analysis results indicated that the bootstrap tests could detect significant differences more than the alternative approaches.

Finally, we developed novel asymptotic-based bootstrap tests for matched pairs with missing values in a single arm. We could confirm that our suggested bootstrap tests are asymptotically correct and robust under heteroscedasticity and/or skewed distributions.

## 5.4 Article IV: Incompletely Observed Nonparametric Factorial Designs with Repeated Measurements: A wild bootstrap approach.

In our last work, we consider the nonparametric statistical model for repeated measures design described in Section 3.2.2. Multivariate analysis of variance (MANOVA) or mixed models is commonly used to analyze multivariate data, requiring complete data and particular assumptions about the underlying parametric distribution, such as continuity or a specific covariance structure, such as compound symmetry. Moreover, if discrete or even ordered categorical data is present, however, mean-based approaches are not applicable. Therefore, the optimal solution is nonparametric rank-based approaches that do not rely on restrictive distributional assumptions and are robust against monotone transformations of the data. Hypotheses are thus no longer formulated in terms of means but rather in terms of distribution functions or relative marginal effects.

However, most available multivariate rank-based approaches have only been developed for complete observations and cannot handle multivariate data with missing values. Only a few approaches are applicable in case of missing values, do not require parametric assumptions, and result in accurate decisions in case of arbitrary covariance structures, skewed distributions, or unbalanced experimental designs. Possible approaches are the two quadratic form tests; rank-based Wald and ANOVA-type test statistics (Brunner et al., 1999; Domhof et al., 2002). However, the Wald-type test is an asymptotically correct test that often requires large sample sizes to obtain accurate test decisions. In contrast, the ANOVA-type test is based on an approximation of its distribution with a scaled  $\chi^2$ -distribution. Since the approximation does not coincide with the ANOVA-type test limiting distribution under  $H_0$ . The ANOVA-type test is not asymptotically valid and results in a more or less conservative behavior for small sample sizes. The aim of the article was to provide asymptotically correct procedures that can handle missing values, do not require parametric assumptions such as continuity of the distribution functions or nonsingular covariance matrices, and can be used with ordinal or ordered categorical data.

We proposed three different quadratic form-type test statistics and improved their small sample behavior by a wild bootstrap procedure, i.e., the Wald-type statistic  $T_W^*$ , the ANOVA-type statistic  $T_A^*$ , and the modified ANOVA-type statistic  $T_M^*$  which are computed as

$$T_W^* = n\hat{\mathbf{p}}^{*T} \mathbf{C}^\top [\mathbf{C}\hat{\mathbf{V}}_n^* \mathbf{C}^\top]^{-1} \mathbf{C}\hat{\mathbf{p}}^*, \quad (5.7)$$

$$T_A^* = \frac{1}{\text{tr}(\mathbf{T}\hat{\mathbf{V}}_n^*)} n\hat{\mathbf{p}}^{*T} \mathbf{T}\hat{\mathbf{p}}^*, \quad (5.8)$$

$$T_M^* = n\hat{\boldsymbol{p}}^{*T}\boldsymbol{C}^\top[\boldsymbol{C}\hat{\boldsymbol{D}}_n^*\boldsymbol{C}^\top]^+\boldsymbol{C}\hat{\boldsymbol{p}}^*, \quad (5.9)$$

where  $\hat{\boldsymbol{p}}^*$  is the bootstrap version of the relative effect estimator vector  $\hat{\boldsymbol{p}}$ ,  $\hat{\boldsymbol{V}}_n^*$  is the bootstrap covariance matrix estimator,  $\boldsymbol{T} = \boldsymbol{C}^\top[\boldsymbol{C}\boldsymbol{C}^\top]^+\boldsymbol{C}$  is a projection matrix, and  $\hat{\boldsymbol{D}}_n^* = \text{diag}(\hat{\boldsymbol{V}}_n^*)_{ii}$ . We proved that the conditional distribution of the Wald, ANOVA, and MATS-type bootstrap statistics  $T_W^*$ ,  $T_A^*$ , and  $T_M^*$  approximate the null distribution of  $T_W$ ,  $T_A$ , and  $T_M$ , respectively. We showed that under  $H_0$ , the wild bootstrap tests are asymptotic level  $\alpha$  tests and are consistent for any fixed alternative. Furthermore, they have the same local power under contiguous alternatives as their original tests (Theorem 4.2 and (IV)).

In addition to the theoretical findings, we have conducted an extensive simulation study to investigate the small sample behavior of the suggested wild bootstrap tests and their asymptotic quadratic form tests counterparts. We considered a two-way layout design with  $a = 2$  independent groups and two different time points  $d \in \{4, 8\}$  underlying discrete and continuous distributions and various covariance structures. We generated missingness within MCAR as well as MAR frameworks. Our simulation findings revealed that the asymptotic Wald-type test exhibits an extremely liberal behavior in all scenarios and under all considered MCAR and MAR mechanisms. Moreover, the asymptotic ANOVA-type test provides a quite accurate type-I error control for large sample sizes. However, under small sample sizes, is sensitive to missing rates, and it shows a liberal behavior for larger missing rates in particular. However, our suggested bootstrap approaches tend to result in rather accurate type-I error rate control under both symmetric and skewed distributions, as well as under the MCAR and MAR mechanisms. Less strict missing mechanisms do not affect type-I error control, and bootstrap tests are robust even when there are many missing data. However, the bootstrapped MATS' type-I error control behavior is dependent on the hypothesis of interest.

In addition, we considered two real-life data examples. We considered the fluvoxamine trial, which consisted of 315 patients with psychiatric symptoms. The patients were examined every two weeks over six weeks of treatment ( $d = 3$ ). Scores for both side effect and therapeutic effect scales were recorded. We also considered the skin disorder trial, which aimed to determine the severity of the skin problem over time and evaluate the efficacy of two continuous therapy treatments, drugs, and placebo. Patients were measured prior to therapy to assess the severity of their skin disorder (moderate or severe). The treatment outcome was assessed using a five-point ordinal response scale at three follow-up sessions. However, in both data examples, several patients missed attending some sessions, resulting in a large number of missing values. Our analysis revealed that all tests show a significant difference between the therapeutic effect scores of the three sessions of the fluvoxamine trial (the same for the side effect scores). Furthermore, all approaches imply that patients' clinical

outcomes improve significantly with time and that this progression differs significantly between the two treatment groups (drug and placebo).

To summarize, we have developed three asymptotically correct bootstrap procedures that can handle missing values, allow for singular covariance matrices, and apply for ordinal or ordered categorical data. Besides, we confirmed their applicability to data with small sample sizes. Finally, we recommend our proposed bootstrap ANOVA-type test that has the best overall type-I error control and good power behavior among all considered tests.

## 6 Outlook: General MANOVA with Missing Data

For repeated measures designs, we have looked at a semiparametric model of matched pairs data with missing values in both arms ((I), (II)) and a single arm (III) and a nonparametric model for matched pairs with missing values in both arms (II). In (IV), we also looked at a nonparametric factorial design with repeated measures data. A semiparametric model for multivariate data with missing values is another scenario in this setting. We provide a brief summary of the related working paper by Amro et al. in this chapter.

We consider a general model with  $a$  independent and potentially unbalanced treatment groups and  $d$ -variate measurements given by independent random vectors

$$\mathbf{X}_{ik} = \boldsymbol{\mu} + \boldsymbol{\epsilon}_{ik}, \quad i = 1, \dots, a; \quad k = 1, \dots, n_i,$$

where  $\boldsymbol{\mu} = [\boldsymbol{\mu}_1^\top, \dots, \boldsymbol{\mu}_a^\top]^\top$ ,  $\boldsymbol{\mu}_i = [\mu_i^{(1)}, \dots, \mu_i^{(d)}]^\top$ . For fixed  $i$ , the error terms  $\epsilon_{i1}, \dots, \epsilon_{in_i}$  are assumed to be independent and identically distributed  $d$ -dimensional random vectors with:

- $E(\boldsymbol{\epsilon}_{ik}) = \mathbf{0}$ ,  $i = 1, \dots, a; \quad k = 1, \dots, n_i$ ,
- $Cov(\boldsymbol{\epsilon}_{ik}) = \boldsymbol{\Sigma}_i \geq \mathbf{0}$ ,  $i = 1, \dots, a; \quad k = 1, \dots, n_i$ ,
- $E(\|\boldsymbol{\epsilon}_{ik}\|^4) < \infty$ ,  $i = 1, \dots, a; \quad k = 1, \dots, n_i$ .

Thus, the only distributional assumption is the existence of finite second moments, i.e.,  $\boldsymbol{\Sigma}_i \geq \mathbf{0}$ . In order to include the case of missing values, we follow the notation of Brunner et al. (1999) and let

$$\lambda_{ijk} = \begin{cases} 1, & \text{if } X_{ijk} \text{ is observed} \\ 0, & \text{if } X_{ijk} \text{ is non-observed} \end{cases} \quad i = 1, \dots, a; \quad j = 1, \dots, d; \quad k = 1, \dots, n_i. \quad (6.1)$$

Moreover, let  $n = \sum_{i=1}^a n_i$  denote the total number of subjects and let  $N = \sum_{i=1}^a \sum_{j=1}^d \sum_{k=1}^{n_i} \lambda_{ijk}$  denote the total number of observations.

To formulate the null hypothesis in this semiparametric setup, let  $\mathbf{C}$  denote a contrast matrix, i.e.,  $\mathbf{C}\mathbf{1} = \mathbf{0}$  where  $\mathbf{1} = [1, \dots, 1]^\top$  and  $\mathbf{0} = [0, \dots, 0]^\top$ . Then, the null hypotheses are formulated by  $H_0 : \{\mathbf{C}\boldsymbol{\mu} = \mathbf{0}\}$ . This framework covers different

factorial designs. We derive asymptotic theory under the following sample size assumption and missing values:

$$\frac{\lambda_{ij.}}{n} \rightarrow \kappa_i \in (0, 1) \quad i = 1, \dots, a; \quad j = 1, \dots, d, \quad \text{as } \min\{\lambda_{ij.}\} \rightarrow \infty.$$

An estimator of the mean vector  $\mu$  is given by

$$\bar{\mathbf{X}} = [\bar{\mathbf{X}}_1^\top, \dots, \bar{\mathbf{X}}_a^\top],$$

where  $\bar{\mathbf{X}}_i = \frac{1}{\lambda_{ij.}} \sum_{k=1}^{n_i} \lambda_{ijk} \mathbf{X}_{ik}$ .

The covariance matrix of  $\sqrt{n}\bar{\mathbf{X}}$  is given by:

$$\mathbf{V}_n = \text{Cov}(\sqrt{n}\bar{\mathbf{X}}) = \text{diag}\left(\frac{n}{n_i} \mathbf{V}_i, 1 \leq i \leq a\right),$$

where the covariance matrix  $\mathbf{V}_n$  is estimated by

$$\hat{\mathbf{V}}_n = \text{diag}\left(\frac{n}{n_i} \hat{\mathbf{V}}_i, i = 1, \dots, a\right),$$

where  $\hat{\mathbf{V}}_i = [\hat{v}_i(j, j')]$  with diagonal and off-diagonal elements  $\hat{v}_i(j, j)$  and  $\hat{v}_i(j, j')$ , respectively, defined as

$$\hat{v}_i(j, j) = \frac{n_i}{\lambda_{ij.}(\lambda_{ij.} - 1)} \sum_{k=1}^{n_i} \lambda_{ijk} [X_{ijk} - \bar{X}_{ij.}]^2,$$

$$\hat{v}_i(j, j') = \frac{n_i}{(\lambda_{ij.} - 1)(\lambda_{ij'.} - 1) + \Lambda_{i,jj'} - 1} \sum_{k=1}^{n_i} \lambda_{ijk} \lambda_{ij'k} [(X_{ijk} - \bar{X}_{ij.})(X_{ij'k} - \bar{X}_{ij'.})].$$

Most commonly used test statistics for multivariate data are the Wald-type statistic (WTS), ANOVA-type statistic (ATS), and modified ANOVA-type statistic; in our setting, we define them as follows:

### Wald-type Statistic

$$T_W = [\mathbf{C}\bar{\mathbf{X}}_n]^\top [\mathbf{C}\hat{\mathbf{V}}_n\mathbf{C}^\top]^{-1} [\mathbf{C}\bar{\mathbf{X}}_n],$$

### ANOVA-type Statistic

$$T_A = \frac{[\mathbf{C}\bar{\mathbf{X}}_n]^\top [\mathbf{C}\bar{\mathbf{X}}_n]}{\text{tr}(\mathbf{C}\hat{\mathbf{V}}_n\mathbf{C}^\top)},$$



### Modified ANOVA-type statistic

$$T_M = [\mathbf{C}\bar{\mathbf{X}}_n]^\top \hat{\mathbf{D}}[\mathbf{C}\bar{\mathbf{X}}_n],$$

where  $\hat{\mathbf{D}} = \text{diag}([\mathbf{C}\hat{\mathbf{V}}_n\mathbf{C}^\top]_{ii}^+)$  and  $[\cdot]^+$  is the Moore-Penrose inverse.

We want to improve the small sample behavior of the above methods using a parametric bootstrap procedure which can be described as follows:

1. For the given incomplete paired data, calculate the observed test statistic, say  $T$ .
2. Estimate the covariance matrix of each group  $\mathbf{V}_{i,n_i}$  by  $\hat{\mathbf{V}}_{i,n_i}$ ,  $i = 1, \dots, a$ .
3. Generate a bootstrap sample  $\mathbf{X}_{i1}^{*T}, \dots, \mathbf{X}_{in_i}^{*T}$  from  $N(\mathbf{0}, \hat{\mathbf{V}}_{i,n_i})$ .
4. Calculate the value of the test statistic for the bootstrapped sample  $T^*$ .
5. Use its conditional distribution to calculate critical values.

We investigated the suggested bootstrap methods' type-I error control in a simulation study. We considered several settings, including various distributions with different covariance structures and varying sample sizes. The bootstrap approaches showed an accurate type-I error control under almost all settings. Our tests seemed to be robust against heteroscedasticity or skewed distributions in most considered situations.

In summary, we proposed bootstrap procedures to multivariate data under a semi-parametric setup. We have investigated the behavior of the suggested method under the MCAR mechanism, and we are willing to study the bootstrap tests behavior under some MAR and MNAR scenarios. We are currently doing the theoretical investigations and will compile the results in paper form. Also, we will implement the suggested methods in the R package MissPair.

## 7 Conclusions and Outlook

The present dissertation was concerned with employing resampling techniques to develop statistical procedures for repeated measures data with missing values, particularly for data with small sample sizes. It included four articles ((I) - (IV)) that presented the proposed resampling approaches, which were developed under four different statistical models, including semiparametric models where inference is based on means and very general nonparametric models where rank statistics are used which make it applicable for ordinal or ordered categorical data as well as continuous data. Therein, we considered different repeated measures designs and missingness patterns; matched pairs design with missingness in both arms, matched pairs design with missingness in a single arm, and incomplete factorial design with repeated measurements. The latter design allows testing for main and interaction effects between different factors.

For all our proposed approaches, we proved their asymptotic validity, demonstrated their applicability to data with small sample sizes using Monte Carlo simulations and real data examples from various life science research areas. In all considered scenarios, our novel resampling methods improved the small sample behavior of their corresponding asymptotic test statistics. In particular, for the semiparametric model of matched pairs design with missingness in both arms, we proposed a novel weighted permutation test in (I) which is asymptotically valid and even finitely exact if certain invariance properties are met. In addition to this, its favorable small sample properties make it recommendable in practice. Further, we developed a novel multiplication combination approach in (II) that is flexible and can be used to test different hypotheses in nonparametric as well as semiparametric and parametric models for matched pairs design with missingness in both arms. When combined with permutation tests, our MCT procedures resulted in asymptotically valid tests which are even finitely exact under certain invariance properties. Particularly, our proposed MCT enhanced the type-I error control of the existing approaches for the nonparametric Behrens-Fisher problem  $H_0^p : \{p = 1/2\}$ .

Furthermore, for the special missing pattern, matched pairs with missing values in one arm only, we proposed under a semiparametric model adequate and asymptotically correct testing procedures based on wild bootstrap versions of WTS, ATS, MATS, and Little, which overcome the existing approaches (III). Moreover, for the incompletely observed nonparametric factorial designs with repeated measurements, we proposed wild bootstrap versions of the quadratic forms: rank based Wald, ANOVA, and

MATS-type statistics (IV). Based upon the simulation results, the best test behavior was exhibited by the wild bootstrap version of the ANOVA-type statistic and is recommended.

All of our proposed methods in this dissertation are not based on homoscedasticity, multivariate normality, or balanced experimental designs assumptions. They use all the observed information. They also lead to valid inferences in case of arbitrary covariance structures, heteroscedasticity, skewed distributions, or unbalanced designs. Furthermore, our proposed tests showed a robust behavior under fairly large amounts of missing observations. Thus, our proposed methods should be broadly applicable in various practical fields.

Since all our proposed approaches are only applicable for repeated measures data, we are currently developing bootstrap procedures for general MANOVA with missing values under the semiparametric model defined in Chapter 6. We also plan to extend the results of Dobler et al. (2020) for the general MANOVA settings where hypotheses are formulated in terms of unweighted nonparametric effects to the situation with missing values. Additionally, future work will include developing randomization procedures that can handle matched pairs data with missingness in just one arm under a purely nonparametric model. Moreover, all of the methods suggested in this dissertation are based on the MCAR assumption. We want to extend our work in the future by developing resampling procedures for handling incomplete data within the MAR framework.

Finally, we developed an R package with a web-based Shiny app that contains all our proposed resampling methods for matched pairs data with missing values in single or both arms that will soon be available on the CRAN repository. In addition, our proposed wild bootstrap procedures for the incomplete factorial designs with repeated measurements will be implemented in the R package `nparLD` (Noguchi et al., 2012). This will allow researchers from several fields to access our developed statistical methods freely and easily to apply on their own data sets. More possible extensions and topics for future research are given in the last section of each of our four articles.

# Bibliography

- Akritis, M. G. (2011). Nonparametric models for ANOVA and ANCOVA designs. In M. Lovric (Ed.), *International encyclopedia of statistical science* (pp. 964–968). Springer.
- Akritis, M. G., and Arnold, S. F. (1994). Fully nonparametric hypotheses for factorial designs I: Multivariate repeated measures designs. *Journal of the American Statistical Association*, *89*(425), 336–343.
- Akritis, M. G., and Brunner, E. (1997). A unified approach to rank tests for mixed models. *Journal of Statistical Planning and Inference*, *61*(2), 249–277.
- Allison, P. D. (2001). *Missing data*. Sage publications.
- Amro, L., and Pauly, M. (2017). Permuting incomplete paired data: A novel exact and asymptotic correct randomization test. *Journal of Statistical Computation and Simulation*, *87*(6), 1148–1159.
- Arnau, J., Bono, R., Blanca, M. J., and Bendayan, R. (2012). Using the linear mixed model to analyze nonnormal data distributions in longitudinal designs. *Behavior Research Methods*, *44*(4), 1224–1238.
- Bell, M. L., Fiero, M., Horton, N. J., and Hsu, C.-H. (2014). Handling missing data in RCTs; a review of the top medical journals. *BMC Medical Research Methodology*, *14*(1), 1–8.
- Bhoj, D. S. (1978). Testing equality of means of correlated variates with missing observations on both responses. *Biometrika*, *65*(1), 225–228.
- Breitung, J., and Gourieroux, C. (1997). Rank tests for unit roots. *Journal of Econometrics*, *81*(1), 7–27.
- Brunner, E., Bathke, A. C., and Konietzschke, F. (2018). *Rank and pseudo-rank procedures for independent observations in factorial designs*. Springer Series in Statistics.
- Brunner, E., Konietzschke, F., Pauly, M., and Puri, M. L. (2017). Rank-based procedures in factorial designs: Hypotheses about non-parametric treatment effects. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *79*(5), 1463–1485.
- Brunner, E., and Munzel, U. (2000). The nonparametric Behrens-Fisher problem: Asymptotic theory and a small-sample approximation. *Biometrical Journal*, *42*(1), 17–25.

- Brunner, E., Munzel, U., Puri, M. L. et al. (1999). Rank-score tests in factorial designs with repeated measures. *Journal of Multivariate Analysis*, 70(2), 286–317.
- Brunner, E., and Puri, M. L. (1996). 19 nonparametric methods in design and analysis of experiments. *Handbook of Statistics*, 13, 631–703.
- Brunner, E., and Puri, M. L. (2001). Nonparametric methods in factorial designs. *Statistical Papers*, 42(1), 1–52.
- Clark, K., Vendt, B., Smith, K., Freymann, J., Kirby, J., Koppel, P., Moore, S., Phillips, S., Maffitt, D., Pringle, M. et al. (2013). The cancer imaging archive (TCIA): Maintaining and operating a public information repository. *Journal of digital imaging*, 26(6), 1045–1057.
- Davidson, R., and Flachaire, E. (2008). The wild bootstrap, tamed at last. *Journal of Econometrics*, 146(1), 162–169.
- Davis, C. S. (2002). *Statistical methods for the analysis of repeated measurements*. Springer.
- Dobler, D., Friedrich, S., and Pauly, M. (2020). Nonparametric MANOVA in meaningful effects. *Annals of the Institute of Statistical Mathematics*, 72(4), 997–1022.
- Domhof, S., Brunner, E., and Osgood, D. W. (2002). Rank procedures for repeated measures with missing values. *Sociological Methods and Research*, 30(3), 367–393.
- Efron, B. (1979). Computers and the theory of statistics: Thinking the unthinkable. *SIAM Review*, 21(4), 460–480.
- Efron, B., and Tibshirani, R. J. (1994). *An introduction to the bootstrap*. CRC press.
- Einsporn, R. L., and Habtzghi, D. (2013). Combining paired and two-sample data using a permutation test. *Journal of Data Science*, 11(4), 767–779.
- Ekbohm, G. (1976). On comparing means in the paired case with incomplete data on both responses. *Biometrika*, 63(2), 299–304.
- Enders, C. K. (2010). *Applied missing data analysis*. Guilford press.
- Fisher, R. A. (1935). *Design of experiments*. Edinburgh; London: Oliver; Boyd.
- Fligner, M. A., and Policello, G. E. (1981). Robust rank procedures for the behrens-fisher problem. *Journal of the American Statistical Association*, 76(373), 162–168.
- Fong, Y., Huang, Y., Lemos, M. P., and Mcelrath, M. J. (2018). Rank-based two-sample tests for paired data with missing values. *Biostatistics*, 19(3), 281–294.
- Friedrich, S., Brunner, E., and Pauly, M. (2017a). Permuting longitudinal data in spite of the dependencies. *Journal of Multivariate Analysis*, 153, 255–265.

- Friedrich, S., Konietzschke, F., and Pauly, M. (2017b). A wild bootstrap approach for nonparametric repeated measurements. *Computational Statistics and Data Analysis*, 113, 38–52.
- Friedrich, S., and Pauly, M. (2018). MATS: Inference for potentially singular and heteroscedastic MANOVA. *Journal of Multivariate Analysis*, 165, 166–179.
- Good, P. (2005). *Permutation, parametric and bootstrap tests of hypotheses*. Springer.
- Good, P. (2006). *Resampling methods*. Springer.
- Guo, B., and Yuan, Y. (2017). A comparative review of methods for comparing means using partially paired data. *Statistical Methods in Medical Research*, 26(3), 1323–1340.
- Hettmansperger, T. P., and McKean, J. W. (2010). *Robust nonparametric statistical methods*. CRC Press.
- Jansen, I., Molenberghs, G., Aerts, M., Thijs, H., and Van Steen, K. (2003). A local influence approach applied to binary data from a psychiatric study. *Biometrics*, 59(2), 410–419.
- Janssen, A. (1997). Studentized permutation tests for non-iid hypotheses and the generalized behrens-fisher problem. *Statistics and Probability Letters*, 36(1), 9–21.
- Janssen, A. (1999). Testing nonparametric statistical functionals with applications to rank tests. *Journal of Statistical Planning and Inference*, 81(1), 71–93.
- Janssen, A. (2005). Resampling student's-t-type statistics. *Annals of the Institute of Statistical Mathematics*, 57(3), 507–529.
- Janssen, A., and Pauls, T. (2003). How do bootstrap and permutation tests work? *The Annals of Statistics*, 31(3), 768–806.
- Kim, B. S., Kim, I., Lee, S., Kim, S., Rha, S. Y., and Chung, H. C. (2005). Statistical methods of translating microarray data into clinically relevant diagnostic information in colorectal cancer. *Bioinformatics*, 21(4), 517–528.
- Koboldt, D., Fulton, R., McLellan, M., Schmidt, H., Kalicki-Veizer, J., McMichael, J., Fulton, L., Dooling, D., Ding, L., Mardis, E. et al. (2012). Comprehensive molecular portraits of human breast tumours. *Nature*, 490(7418), 61–70.
- Konietzschke, F., Bathke, A. C., Harrar, S. W., and Pauly, M. (2015). Parametric and nonparametric bootstrap methods for general MANOVA. *Journal of Multivariate Analysis*, 140, 291–301.
- Konietzschke, F., Harrar, S. W., Lange, K., and Brunner, E. (2012). Ranking procedures for matched pairs with missing data—asymptotic theory and a small sample approximation. *Computational Statistics and Data Analysis*, 56(5), 1090–1102.

- Konietschke, F., and Pauly, M. (2014). Bootstrapping and permuting paired  $t$ -test type statistics. *Statistics and Computing*, 24(3), 283–296.
- Kostecki-Dillon, T., Monette, G., and Wong, P. (2018). Pine trees, comas and migraines. *Newsletter. York University Institute for Social Research*, 14, 1–4.
- Lin, P.-E., and Stivers, L. E. (1974). On difference of means with incomplete data. *Biometrika*, 61(2), 325–334.
- Little, R. J. (1976). Inference about means from incomplete multivariate data. *Biometrika*, 63(3), 593–604.
- Little, R. J., and Rubin, D. B. (2019). *Statistical analysis with missing data*. John Wiley Sons.
- Looney, S. W., and Jones, P. W. (2003). A method for comparing two normal means using combined samples of correlated and uncorrelated data. *Statistics in Medicine*, 22(9), 1601–1610.
- Mammen, E. (1992). Bootstrap, wild bootstrap, and asymptotic normality. *Probability Theory and Related Fields*, 93(4), 439–455.
- Maritz, J. S. (1995). A permutation paired test allowing for missing values. *Australian Journal of Statistics*, 37(2), 153–159.
- Munzel, U. (1999). Nonparametric methods for paired samples. *Statistica Neerlandica*, 53(3), 277–286.
- Munzel, U., and Brunner, E. (2000). Nonparametric methods in multivariate factorial designs. *Journal of Statistical Planning and Inference*, 88(1), 117–132.
- Noguchi, K., Gel, Y. R., Brunner, E., and Konietschke, F. (2012). nparLD: An R software package for the nonparametric analysis of longitudinal data in factorial experiments. *Journal of Statistical software*, 50(12).
- Pauly, M., Brunner, E., and Konietschke, F. (2015). Asymptotic permutation tests in general factorial designs. *Journal of the Royal Statistical Society: Series B: Statistical Methodology*, 461–473.
- Pesarin, F., and Salmaso, L. (2010). *Permutation tests for complex data: Theory, applications and software*. John Wiley and Sons.
- Pesarin, F., and Salmaso, L. (2012). A review and some new results on permutation testing for multivariate problems. *Statistics and Computing*, 22(2), 639–646.
- Peterson, P. L., Baker, E., and McGaw, B. (2010). *International encyclopedia of education*. Elsevier Ltd.
- Pitman, E. J. (1937). Significance tests which may be applied to samples from any populations. *Supplement to the Journal of the Royal Statistical Society*, 4(1), 119–130.

- Qi, Q., Yan, L., and Tian, L. (2019). Testing equality of means in partially paired data with incompleteness in single response. *Statistical Methods in Medical Research*, 28(5), 1508–1522.
- Ramosaj, B., Amro, L., and Pauly, M. (2020). A cautionary tale on using imputation methods for inference in matched-pairs design. *Bioinformatics*, 36(10), 3099–3106.
- Riordan, B. (2012). Northeastern ohio coyote hybridization with wolves. *Honors research project, University of Akron, Akron*.
- Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3), 581–592.
- Salmaso, L. (2015). Combination-based permutation tests: Equipower property and power behavior in presence of correlation. *Communications in Statistics-Theory and Methods*, 44(24), 5225–5239.
- Samawi, H. M., and Vogel, R. (2014). Notes on two sample tests for partially correlated (paired) data. *Journal of Applied Statistics*, 41(1), 109–117.
- Shao, J., and Tu, D. (2012). *The jackknife and bootstrap*. Springer Science and Business Media.
- Xu, J., and Harrar, S. W. (2012). Accurate mean comparisons for paired samples with missing data: An application to a smoking-cessation trial. *Biometrical journal*, 54(2), 281–295.
- Xu, L.-W., Yang, F.-Q., Qin, S. et al. (2013). A parametric bootstrap approach for two-way ANOVA in presence of possible interactions with unequal variances. *Journal of Multivariate Analysis*, 115, 172–180.
- Yu, D., Lim, J., Liang, F., Kim, K., Kim, B. S., and Jang, W. (2012). Permutation test for incomplete paired data with application to cDNA microarray data. *Computational Statistics and Data Analysis*, 56(3), 510–521.



# Part II

## Publications



In this part of the dissertation, you can partially find the original articles on which the dissertation is based. Because of publishing rights, the articles (I), (II), and (IV) could not be included here. Since article (III) is published under the terms of the Creative Commons Attribution License (CC-BY), the final published version is included on the next page.

# Asymptotic-based bootstrap approach for matched pairs with missingness in a single arm

Lubna Amro  | Markus Pauly  | Burim Ramosaj 

Mathematical Statistics and Applications in Industry, Faculty of Statistics, Technical University of Dortmund, Dortmund, Germany

## Correspondence

Lubna Amro, Mathematical Statistics and Applications in Industry, Faculty of Statistics, Technical University of Dortmund, 44227 Dortmund, Germany.  
Email: [lubna.amro@tu-dortmund.de](mailto:lubna.amro@tu-dortmund.de)

## Funding information

Deutsche Forschungsgemeinschaft; Deutscher Akademischer Austauschdienst



This article has earned an open data badge “Reproducible Research” for making publicly available the code necessary to reproduce the reported results. The results reported in this article could fully be reproduced.

## Abstract

The issue of missing values is an arising difficulty when dealing with paired data. Several test procedures are developed in the literature to tackle this problem. Some of them are even robust under deviations and control type-I error quite accurately. However, most of these methods are not applicable when missing values are present only in a single arm. For this case, we provide asymptotic correct resampling tests that are robust under heteroskedasticity and skewed distributions. The tests are based on a meaningful restructuring of all observed information in quadratic form-type test statistics. An extensive simulation study is conducted exemplifying the tests for finite sample sizes under different missingness mechanisms. In addition, illustrative data examples based on real life studies are analyzed.

## KEYWORDS

matched pairs, missing values, parametric bootstrap, quadratic forms

## 1 | INTRODUCTION

Conducting statistical tests on units measured repeatedly requires the consideration of the dependence structure of the resulting random vector. The simplest design is the matched pairs model, where units are measured at two endpoints of the same subject. This design has experienced a large field of application, including industrial and life sciences. In Biomedicine for example, several studies have been focused on identifying genes for up- or downregulated effects in head and neck squamous, prostate, lung, or breast cell carcinoma (Kuriakose et al., 2004; Lapointe et al., 2004; Feng et al., 2008). In common statistical analysis, testing the equality of means in matched pairs design is conducted using the paired  $t$ -test. Even for nonnormal data, the procedure is asymptotically exact, that is, for sufficiently large samples, the test procedure is correctly reflecting type-I error. However, first limitation of the paired  $t$ -test arises when data are only partially observed. Deleting observations with missing values is a suboptimal solution, since variance or mean estimation based only on complete case analysis can be biased leading to incorrect statistical inference. This is especially the case when complete samples are small.

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2020 The Authors. *Biometrical Journal* published by Wiley-VCH GmbH.

To tackle this issue, a simple approach is to impute missing values singly (or multiply) and to carry out statistical tests as if there were no missing values so far (Schafer, 1999; Rubin, 2004; Sterne et al., 2009). However, although leading to good imputation error (Stekhoven & Bühlmann, 2011; Waljee et al., 2013; Ramosaj & Pauly, 2019), such approaches may lead to inflated type-I error rate or remarkably low power in small to moderate sample sizes (Van Buuren, 2018; Ramosaj et al., 2020). Therefore, we do not follow this approach here.

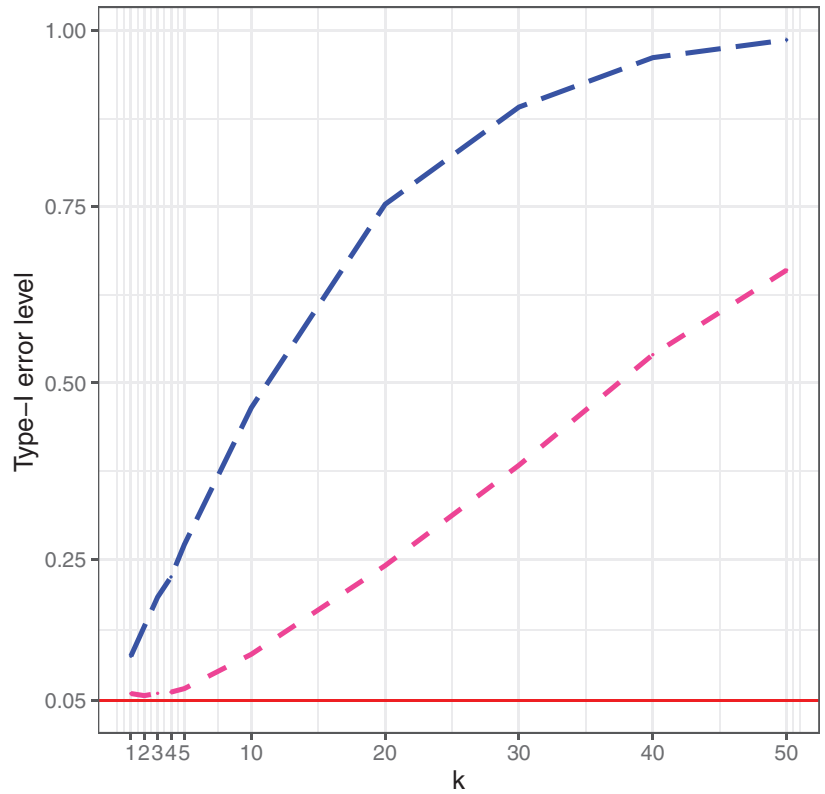
Differing to imputation, several test procedures that (only) use all observed information in the matched pairs design have been proposed in the literature (Mehta & Gurland, 1969; Lin, 1973; Morrison, 1973; Lin & Stivers, 1974; Little, 1976; Ekbohm, 1976; Bhoj, 1978; Looney & Jones, 2003; Kim et al., 2004; Xu & Harrar, 2012; Fuchs et al., 2017; Uddin & Hasan, 2017). These tests, however, rely on specific model assumptions such as symmetry or even bivariate normality, which are hard to verify in practice. Moreover, these procedures are usually nonrobust to deviations and may result in inaccurate decisions caused by possibly conservative or inflated type-I error rates (Samawi & Vogel, 2014; Amro & Pauly, 2017; Amro et al., 2019; Qi et al., 2019; Harrar et al., 2020).

To overcome these problems, the typical recommendation is to use the method based on combining separate results of adequate test statistics for the underlying paired and unpaired portions of the data using either weighted test statistics (Samawi & Vogel, 2014; Amro & Pauly, 2017; Martínez-Cambor et al., 2013), a multiplication combination test (Amro et al., 2019), or combined  $p$ -values (Rempala & Looney, 2006; Samawi et al., 2011; Yu et al., 2012; Kuan & Huang, 2013). However, all these methods are only applicable for matched pairs with missingness in both arms. This is due to their tests construction. Since, they are based upon combining the results of two independent tests for the related paired and unpaired two-sample problem. As independence of these two tests is required, a direct adjustment to handle data with missingness in one arm only is not possible. Thus, these methods cannot be used to analyze data on pathological stage I breast cancer patients from the Cancer Genome Atlas (TCGA) project. This data set consists of observations from 90 patients of which 74 had entries in one component of it, only 16 were complete, see Section 7.1 for details. The question is now how to analyze such data?

In contrast to the above methods, barely any work can be found that is potentially applicable in this special missing pattern, requires no parametric assumptions and also leads to valid inferences in case of heteroskedasticity or skewed distributions. One exception is given by the recent proposals of Qi et al. (2019) who recommended a so-called nonparametric combination test (NCT) and nonparametric  $p$ -value pooling methods (NPM). The NCT is based on merging the results from *Sign test* and *Wilcoxon Mann–Whitney test* while the NPM are based on combining  $p$ -values of the *Wilcoxon signed-rank test* and *Mann–Whitney test*. In situations where these two nonparametric procedures show their efficiency, their proposed combination is indeed tempting. However, neither the *Sign test* is known to be very powerful for metric data nor is the *Mann–Whitney test* known for being robust against heteroskedasticity. In fact, our simulation studies demonstrate that the NCT and Fisher's pooling method (FPM) as an NPM inherit these unsatisfying properties to some extent: under heteroskedasticity and/or skewed distributions, the NCT and FPM tend to not maintain the pre-assigned type-I error level. The degree of variance heterogeneity, skewness, and sample sizes can all affect the type-I error rate control level. An example of the type-I error control of NCT and FPM when heteroskedasticity coincides with a skewed error distribution is displayed in Figure 1. It reveals that, under heteroskedasticity and an exponential distribution, the NCT and FPM type-I error rate functions become surprisingly analogous to the power function where the type-I error rate increases dramatically with an increase in sample sizes.

The aim of this paper is therefore bilateral: First, we aim to provide a statistical test that is capable of treating single-arm missing values in matched pairs which drop the common assumptions such as homoskedasticity and normality, while not losing (partial) information. Second, it should be able to satisfactorily control type-I error while maintaining good power properties. To this end, we propose three different test statistics, analyze their asymptotic behaviors under the null hypothesis and equip them with an asymptotically correct parametric bootstrap procedure for calculating critical values. In doing so, we structured the paper by first introducing the statistical model and the hypothesis of interest. In Section 3, we provide different test statistics of quadratic form-type that either converge to a  $\chi^2$  or a weighted  $\chi^2$ -distribution. Proofs presenting theoretical guarantees of the proposed methods are delivered in the supplement. In Section 4, we introduce a parametric bootstrap technique to calculate critical values and prove its theoretical correctness. Section 5 is devoted to already existing methods for statistical inference in matched pairs with single-arm missingness while in Sections 6 and 7, novel and existing methods are compared based on an extensive simulation study and three real life data examples. The supplement contains additional theoretical details. For notational purposes, we state vectors or matrices in bold and scalars in usual form.

**FIGURE 1** Type-I error simulation results ( $\alpha = .05$ ) of the nonparametric combination test  $T_N$  (—) and the Fisher’s  $p$ -value pooling method  $T_F$  (---) for exponential distribution under correlation factor ( $\rho = .7$ ) and a heteroskedastic setup with variances 1 and 2, respectively, for increasing sample sizes  $k \cdot (n_c, n_u) = (k \cdot 10, k \cdot 30)$  under the MCAR framework



## 2 | STATISTICAL MODEL AND HYPOTHESES

We consider matched pairs given by a sample  $D_n := \{\mathbf{X}_1, \dots, \mathbf{X}_n\}$ , where  $\mathbf{X}_j = [X_{1j}, X_{2j}]^\top \in \mathbb{R}^2$  are i.i.d. random vectors with mean vector  $\mathbb{E}[\mathbf{X}_1] = \boldsymbol{\mu} = [\mu_1, \mu_2]^\top \in \mathbb{R}^2$  and an arbitrary covariance matrix  $0 < \boldsymbol{\Gamma} = \begin{bmatrix} \tilde{\sigma}_1^2 & \rho \tilde{\sigma}_1 \tilde{\sigma}_2 \\ \rho \tilde{\sigma}_1 \tilde{\sigma}_2 & \tilde{\sigma}_2^2 \end{bmatrix} \in \mathbb{R}^{2 \times 2}$ , where  $\tilde{\sigma}_1^2 = \text{var}(X_{11})$ ,  $\tilde{\sigma}_2^2 = \text{var}(X_{21})$  and  $\rho = \text{corr}(X_{11}, X_{21})$ . To incorporate missingness in one arm (says, the second) only denote with  $R_{2j} \in \{0, 1\}$ ,  $j = 1, \dots, n$  the vector whose  $j$ th component indicates whether  $X_{2j}$  is observed ( $R_{2j} = 1$ ) or missing ( $R_{2j} = 0$ ) for  $j = 1, \dots, n$ . Define the composition  $*$  by  $a * 1 = a$  and  $a * 0 = ---$ , for all  $a \in \mathbb{R}$ , then in practice, one observes  $\mathbf{X}^{(o)} := \{\mathbf{X}_j * \mathbf{R}_j\}_{j=1}^n$  where  $\mathbf{R}_j = [1, R_{2j}]^\top \in \mathbb{R}^2$ ,  $j = 1, \dots, n$ , and a “---” entry is interpreted as missing. Hence, our framework has the following form:

$$\underbrace{\begin{bmatrix} X_{11}^{(c)} \\ X_{21}^{(c)} \end{bmatrix}, \dots, \begin{bmatrix} X_{1n_c}^{(c)} \\ X_{2n_c}^{(c)} \end{bmatrix}}_{\mathbf{X}^{(c)}}, \underbrace{\begin{bmatrix} X_{11}^{(i)} \\ --- \end{bmatrix}, \dots, \begin{bmatrix} X_{1n_u}^{(i)} \\ --- \end{bmatrix}}_{\mathbf{X}^{(i)}}. \tag{1}$$

Rubin defines the missing mechanism through a parametric distributional model on  $\mathbf{R} = \{\mathbf{R}_j\}_{j=1}^n$  and classifies their presence through Missing Completely at Random (MCAR), Missing at Random (MAR), and Missing not at Random (MNAR) schemes (Rubin, 2004). In our work, we first assume an MCAR mechanism, in that  $\mathbf{X}^{(c)}$  is independent of  $\mathbf{X}^{(i)}$ . However, we will also study MAR mechanisms in simulations and relate to the supplement for the explicit definition of the missing mechanisms. For notational purposes, let  $I_{n_c}$  denote the index set of  $|I_{n_c}| = n_c$  complete pairs, that is,  $\mathbf{R}_j = [1, 1]^\top$  for all  $j \in I_{n_c}$ . Similarly,  $I_{n_u}$  is the index set of observations with second component missing ( $\mathbf{R}_j = [1, 0]^\top$ ,  $j \in I_{n_u}$ ) and  $|I_{n_u}| = n_u$ . Thus, there are in total  $N = 2n_c + n_u$  observations from  $n = n_c + n_u$  subjects.

In this framework, we would like to use all the available data to test the null hypothesis  $H_0 : \{\mu_1 = \mu_2\}$  of equal means against the alternative  $H_1 : \{\mu_1 \neq \mu_2\}$ .

To construct our test statistics, we first fix estimators of the population means  $\mu_1$ , and  $\mu_2$ . For estimating  $\mu_1$ , we consider two estimators; the sample mean of the first components of the completed data set  $\bar{X}_1^{(c)} = \frac{1}{n_c} \sum_{i=1}^{n_c} X_{1i}^{(c)}$ , and the sample

mean of the first components of the unpaired data  $\bar{X}_{1.}^{(i)} = \frac{1}{n_u} \sum_{j=1}^{n_u} X_{1j}^{(i)}$ . For estimating the population mean  $\mu_2$ , we use the sample mean of the second components of the complete data  $\bar{X}_{2.}^{(c)} = \frac{1}{n_c} \sum_{i=1}^{n_c} X_{2i}^{(c)}$ . Next, we define the normalized vector  $\mathbf{Z}_n$  that aggregates the difference between the mean values  $\boldsymbol{\mu} = [\mu_1, \mu_2]^\top$  and their empirical estimators  $[\bar{X}_{1.}^{(c)}, \bar{X}_{2.}^{(c)}, \bar{X}_{1.}^{(i)}]^\top$

$$\mathbf{Z}_n = \sqrt{n}[\bar{X}_{1.}^{(c)} - \mu_1, \bar{X}_{2.}^{(c)} - \mu_2, \bar{X}_{1.}^{(i)} - \mu_1]^\top \tag{2}$$

and take their correlations into account in the covariance matrix

$$\boldsymbol{\Sigma}_n := \text{cov}(\mathbf{Z}_n) = \begin{bmatrix} (n/n_c)\sigma_1^2 & (n/n_c)\rho\sigma_1\sigma_2 & 0 \\ (n/n_c)\rho\sigma_1\sigma_2 & (n/n_c)\sigma_2^2 & 0 \\ 0 & 0 & (n/n_u)\sigma_1^2 \end{bmatrix},$$

where  $\sigma_1^2 = \text{var}(X_{11}^{(c)}) = \text{var}(X_{11}^{(i)})$ ,  $\sigma_2^2 = \text{var}(X_{21}^{(c)})$ , and  $\rho = \text{corr}(X_{11}^{(c)}, X_{21}^{(c)})$ .

To test the null hypothesis  $H_0 : \{\mu_1 - \mu_2 = 0\}$ , we define the two estimators  $\bar{X}_{1.}^{(c)} - \bar{X}_{2.}^{(c)}$  and  $\bar{X}_{1.}^{(i)} - \bar{X}_{2.}^{(c)}$  for  $\mu_1 - \mu_2$ . Their joined asymptotic behavior under the null hypothesis  $H_0$  is studied below.

**Proposition 1.** Set  $f_A(\mathbf{x}) = \mathbf{A}\mathbf{x}$ , for the matrix  $\mathbf{A} = \begin{bmatrix} 1 & -1 & 0 \\ 0 & -1 & 1 \end{bmatrix} \in \mathbb{R}^{2 \times 3}$ . Then, under the null hypothesis  $H_0$  and the condition that  $\frac{n_c}{n_c+n_u} \rightarrow \kappa_1 \in (0, 1)$  and  $\frac{n_u}{n_c+n_u} \rightarrow \kappa_2 = (1 - \kappa_1) \in (0, 1)$  as  $n \rightarrow \infty$ , the composite statistic

$$f_A \circ \mathbf{Z}_n = \mathbf{A}\mathbf{Z}_n = \sqrt{n}[\bar{X}_{1.}^{(c)} - \bar{X}_{2.}^{(c)}, \bar{X}_{1.}^{(i)} - \bar{X}_{2.}^{(c)}]^\top \tag{3}$$

is asymptotically  $N_2(\mathbf{0}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^\top)$  distributed as  $n \rightarrow \infty$ .

$$\text{Here, } \boldsymbol{\Sigma} = \lim_{n \rightarrow \infty} \boldsymbol{\Sigma}_n = \begin{bmatrix} \kappa_1^{-1}\sigma_1^2 & \kappa_1^{-1}\rho\sigma_1\sigma_2 & 0 \\ \kappa_1^{-1}\rho\sigma_1\sigma_2 & \kappa_1^{-1}\sigma_2^2 & 0 \\ 0 & 0 & \kappa_2^{-1}\sigma_1^2 \end{bmatrix}. \tag{4}$$

### 3 | STATISTICS AND ASYMPTOTICS

In this section, we propose three different quadratic forms for testing  $H_0$ : a Wald-type statistic (WTS), an ANOVA-L2-type statistic (ATS), and a modified ANOVA-type statistic (MATS). To introduce the WTS, denote by  $\mathbf{B}^+$  the Moore–Penrose inverse of a matrix  $\mathbf{B}$ . Then, the WTS is given by

$$T_W = (\mathbf{A}\mathbf{Z}_n)^\top (\mathbf{A}\hat{\boldsymbol{\Sigma}}_n\mathbf{A}^\top)^+ (\mathbf{A}\mathbf{Z}_n), \tag{5}$$

where  $\hat{\boldsymbol{\Sigma}}_n$  is the plug-in sample estimator for  $\boldsymbol{\Sigma}$  given in (4), see the supplement for its explicit form. Thanks to the introduced studentization by  $(\mathbf{A}\hat{\boldsymbol{\Sigma}}_n\mathbf{A}^\top)^+$ , the WTS is asymptotically  $\chi^2_2$ -distributed under the null hypothesis as long as  $\boldsymbol{\Sigma} > \mathbf{0}$  as proved in the supplement.

Similar WTS versions are also studied in the context of heteroskedastic ANOVA or MANOVA (Krishnamoorthy & Lu, 2010; Xu et al., 2013; Konietzschke et al., 2015; Friedrich & Pauly, 2018). From these settings, it is known that the convergence to its limiting  $\chi^2$ -distribution is rather slow and large sample sizes are required to obtain adequate results (Vallejo et al., 2010; Konietzschke et al., 2015; Smaga, 2017), which leads to several refinements regarding bootstrapping for the calculations of critical values (see Section 4) or other structures of test statistics. In particular Brunner (2001) proposed an alternative quadratic form by deleting the estimated covariance matrix  $\hat{\boldsymbol{\Sigma}}_n$  involved in the computation of the WTS. Here, we erase the Moore–Penrose inverse term from the WTS resulting in the following ATS:

$$T_A = \frac{1}{\text{tr}(\mathbf{A}\hat{\boldsymbol{\Sigma}}_n\mathbf{A}^\top)} (\mathbf{A}\mathbf{Z}_n)^\top (\mathbf{A}\mathbf{Z}_n). \tag{6}$$

The ATS has the advantage of being applicable in case of singular covariance matrices ( $|\hat{\Sigma}_n| = 0$ ). However, it has the drawback of not being asymptotically distribution-free under the null hypothesis, see the supplement for details.

Another possible test statistic would be the MATS that was developed by Friedrich & Pauly (2018) for MANOVA models. The authors could provide preferable simulation results regarding its power behavior and type-I error control while delivering theoretical guarantees for its validity. Hence, we consider a MATS (with a slight modification) in our design, too. Here, it is given by

$$T_M = (\mathbf{AZ}_n)^\top \hat{\mathbf{D}}_n (\mathbf{AZ}_n), \tag{7}$$

where  $\hat{\mathbf{D}}_n = \text{diag}((\mathbf{A}\hat{\Sigma}_n\mathbf{A}^\top)_{ii}^+)$ .

Similar to the ATS, the MATS is also not distribution-free under  $H_0$ , see the supplement for the explicit form of its limiting distribution. Thus, we cannot directly calculate critical values for  $T_A$  and  $T_M$ , respectively. In addition, the  $\chi^2_2$ -approximation to  $T_W$  is rather slow. To this end, we develop adequate and asymptotically correct testing procedures based on bootstrap versions of  $T_W$ ,  $T_A$ , and  $T_M$  in the subsequent section.

#### 4 | PARAMETRIC BOOTSTRAPPING

To estimate critical values, we apply an asymptotic model-based bootstrap approach which has, for example, been applied in the context of (M)ANOVA factorial designs (Konietschke et al., 2015; Friedrich & Pauly, 2018). To this end, we first generate parametric bootstrap variables as

$$\mathbf{X}_j^* = \begin{bmatrix} X_{1j}^* \\ X_{2j}^* \end{bmatrix} \stackrel{\text{i.i.d.}}{\sim} N(0, \hat{\Gamma}), j = 1, \dots, n. \tag{8}$$

Here,  $\hat{\Gamma} = \begin{bmatrix} \hat{\sigma}_1^2 & \hat{\rho}\hat{\sigma}_1\hat{\sigma}_2 \\ \hat{\rho}\hat{\sigma}_1\hat{\sigma}_2 & \hat{\sigma}_2^2 \end{bmatrix}$  is the empirical covariance matrix estimator, that is,  $\hat{\sigma}_i^2$  denotes the sample variance calculated on all observations in component  $i$  and  $\hat{\rho}$  is the sample correlation obtained from  $\mathbf{X}^{(c)}$ . The idea is to reflect the original covariance structure to obtain more accurate finite sample approximation. Next, we generate missing values under the MCAR scheme by randomly inserting them to the second component of the bivariate vector  $\mathbf{X}_j^*$  until a fixed amount of missing values of size  $n_u$  is achieved. This results into the following bootstrapped data set:

$$\underbrace{\begin{bmatrix} X_{11}^{*(c)} \\ X_{21}^{*(c)} \end{bmatrix}, \dots, \begin{bmatrix} X_{1n_c}^{*(c)} \\ X_{2n_c}^{*(c)} \end{bmatrix}}_{\mathbf{X}^{*(c)}}, \underbrace{\begin{bmatrix} X_{11}^{*(i)} \\ \text{---} \end{bmatrix}, \dots, \begin{bmatrix} X_{1n_u}^{*(i)} \\ \text{---} \end{bmatrix}}_{\mathbf{X}^{*(i)}} \tag{9}$$

and the combined vector  $(f \circ \mathbf{Z}_n)^* = \mathbf{AZ}_n^* = \sqrt{n}(\bar{X}_{1.}^{*(c)} - \bar{X}_{2.}^{*(c)}, \bar{X}_{1.}^{*(i)} - \bar{X}_{2.}^{*(c)})$ . From this, the bootstrapped versions of the quadratic forms, that is, the WTS  $T_W^*$ , the ATS  $T_A^*$ , and the MATS  $T_M^*$  are computed:

$$T_W^* = (\mathbf{AZ}_n^*)^\top (\mathbf{A}\hat{\Sigma}_n^*\mathbf{A}^\top)^+ (\mathbf{AZ}_n^*), \tag{10}$$

$$T_A^* = \frac{1}{\text{tr}(\mathbf{A}\hat{\Sigma}_n^*\mathbf{A}^\top)} (\mathbf{AZ}_n^*)^\top (\mathbf{AZ}_n^*), \tag{11}$$

$$T_M^* = (\mathbf{AZ}_n^*)^\top \hat{\mathbf{D}}_n^* (\mathbf{AZ}_n^*), \tag{12}$$

where  $\hat{\Sigma}_n^* = \hat{\Sigma}_n(\mathbf{X}^{*(c)}, \mathbf{X}^{*(i)})$  and  $\hat{\mathbf{D}}_n^* = \text{diag}((\mathbf{A}\hat{\Sigma}_n^*\mathbf{A}^\top)_{ii}^+)$ .

It is proven in the supplement that all three bootstrapped test statistics approximate the null distribution of the respective test statistic.



To analyze their finite sample performance, we below conduct extensive simulations (Section 6). Before that, we will first discuss other possible candidates from the literature that should or should not be included in our simulation study.

### 5 | COMPARISON WITH EXISTING MODELS

We briefly review the existing literature on methods that can deal with the case of matched pairs with missing values in one arm only. As outlined in the introduction, there only exists a few which we can summarize as follows:

- (a) Simple methods such as: using the paired  $t$ -test while excluding the unpaired data OR using the independent  $t$ -test while ignoring the covariance structure of the data.
- (b) Tests based on modified maximum likelihood estimators (Morrison, 1973; Ekbohm, 1976; Little, 1976).
- (c) Tests based on simple mean difference estimators (Mehta & Gurland, 1969, 1973; Lin, 1973; Ekbohm, 1976).
- (d)  $p$ -Values pooling methods (Qi et al., 2019).
- (e) Weighted linear and nonlinear combination tests (Pesarin & Salmaso, 2010; Qi et al., 2019).

However, none of the methods is free from distributional assumptions and at the same time robust against deviations such as heteroskedasticity and skewed distributions. In particular, the recent paper by Qi et al. (2019) already included a simulation study to compare several of the tests mentioned in (a)–(e). As a conclusion, they recommended a so-called NCT and  $p$ -value pooling methods.

They investigated in their paper two ways of combining the  $p$ -values; a weighted inverse normal method proposed by Hartung (1999) and an FPM suggested by Brown (1975), Kost & McDermott (2002), and Hou (2005). Due to their quite similar behavior, we only include the FPM and the NCT into our simulation study. As additional competitor for these two and the bootstrap procedures proposed in Section 4, we choose the test of Little (1976). The latter assumes that the data follow a bivariate normal distribution and the test statistic is given by

$$T_L = \frac{\bar{X}_1 - \bar{X}_2^{(c)} - \frac{\hat{\rho}\hat{\sigma}_1^{(c)}\hat{\sigma}_2^{(c)}}{(\hat{\sigma}_1^{(c)})^2}(\bar{X}_1 - \bar{X}_1^{(c)})}{\hat{\sigma}_L}, \tag{13}$$

where  $\bar{X}_1 := 1/n(n_c\bar{X}_1^{(c)} + n_u\bar{X}_1^{(i)})$  and  $\hat{\sigma}_1^{(c)}$  is the empirical standard deviation of  $\{X_{11}^{(c)}, \dots, X_{1n_c}^{(c)}\}$ . Moreover, setting  $\hat{\sigma}_{22.1}^2 = \hat{\sigma}_2^2 - (\hat{\rho}\hat{\sigma}_1^{(c)}\hat{\sigma}_2^{(c)}/(\hat{\sigma}_1^{(c)})^2)$  and  $\hat{\sigma}_X^2 = \hat{\sigma}_{22.1}^2 + \frac{(\hat{\rho}\hat{\sigma}_1^{(c)}\hat{\sigma}_2^{(c)})^2}{(\hat{\sigma}_1^{(c)})^4}\hat{\sigma}_1^4$ , the denominator is given by Little (1976)

$$\hat{\sigma}_L^2 = \frac{\hat{\sigma}_X^2}{n} + \left(\frac{1}{n_c} - \frac{1}{n}\right) \frac{n_c - 2}{n_c - 3} \hat{\sigma}_{22.1}^2 - \frac{2}{n} \frac{\hat{\rho}\hat{\sigma}_1^{(c)}\hat{\sigma}_2^{(c)}}{(\hat{\sigma}_1^{(c)})^2} \hat{\sigma}_1^2 + \frac{\hat{\sigma}_1^2}{n}. \tag{14}$$

The exact distribution of  $T_L$  is rather complicated and Little suggests to approximate it by a  $t$ -reference distribution with  $n_c - 1$  degrees of freedom, that is, the test is given by  $\varphi_L := \mathbb{1}\{|T_L| > t_{n_c-1, 1-\alpha/2}\}$  for some level  $\alpha \in (0, 1)$ . To enhance its small sample properties (see the simulation results for  $\varphi_L$  given in the supplement for details), a parametric bootstrap version of the Little test is studied as well. Similar to  $\varphi_L$ , the resulting Little bootstrap test,  $\varphi_L^* := \mathbb{1}\{|T_L| > c_L^*\}$  is asymptotically correct. Here,  $c_L^*$  denotes the conditional  $(1 - \alpha)$ -quantile of the parametric bootstrap distribution of  $T_L$ .

In addition, the NCT proposed by Qi et al. (2019), is based upon a linear combination of the sign and the Wilcoxon Mann–Whitney test statistics:

$$T_N = T_s + T_m, \tag{15}$$

where  $T_s = \frac{1}{n_c} \sum_{i=1}^{n_c} \phi(X_{1i}^{(c)}, X_{2i}^{(c)})$  and  $T_m = \frac{1}{n_c n_u} \sum_{j=1}^{n_u} \sum_{k=1}^{n_c} \phi(X_{1j}^{(i)}, X_{2k}^{(c)})$  with  $\phi(X_1, X_2) = \begin{cases} 1 & \text{if } X > Y, \\ 1/2 & \text{if } X = Y, \\ 0 & \text{otherwise.} \end{cases}$

It is proposed to approximate the null distribution of  $T_N$  by a normal distribution with mean 1 and variance estimated by  $\widehat{\text{var}}(T_N) = \frac{1}{n_c} + \frac{n_c+n_u+1}{12n_c n_u} + \widehat{\text{cov}}(T_s, T_m)$ , where

$$\widehat{\text{cov}}(T_s, T_m) = \frac{1}{n_c^2 n_u} \sum_{i=1}^{n_c} \sum_{j=1}^{n_u} \mathbb{1}\{X_{1i}^{(c)} > X_{2i}^{(c)}, X_{1j}^{(i)} > X_{2j}^{(c)}\} - \frac{1}{n_c} T_s T_m.$$

Moreover, the NPM proposed by Qi et al. (2019) based upon Fisher’s pooling approach is based upon combining the dependent  $p$ -values of the Wilcoxon signed-rank test  $P_p$  and Mann–Whitney U test  $P_{up}$ . The test statistic is given by

$$T_F = -2\lambda_1 \log(P_p) - 2\lambda_2 \log(P_{up}), \tag{16}$$

where  $\lambda_1$  and  $\lambda_2$  are weights. It was shown that  $T_F$  follows asymptotically a scaled  $c\chi_f^2$ -distribution with  $c = \frac{\text{var}(T_F)}{2E(T_F)}$  and  $f = \frac{2[E(T_F)]^2}{\text{var}(T_F)}$ . Moreover, the mean and variance of  $T_F$  are  $E(T_F) = 2(\lambda_1 + \lambda_2)$ ,  $\text{var}(T_F) = 4(\lambda_1^2 + \lambda_2^2) + 2\lambda_1\lambda_2\eta$ , and  $\eta = \text{Cov}(-2\log(P_p), -2\log(P_{up}))$  Qi et al. (2019). suggested to estimate  $\eta$  by nonparametric bootstrapping to obtain estimates  $\hat{c}$  and  $\hat{f}$  for  $c$  and  $f$ , respectively. Therefore, the null distribution of  $T_F$  can be asymptotically approximated by  $\hat{c}\chi_{\hat{f}}^2$ . In previous simulation studies by Qi et al. (2019), the considered choices of the weights  $\lambda_1$  and  $\lambda_2$  had almost invariant impact on the behavior of FPM. Similar to Qi et al. (2019), we therefore consider the following weights:  $\lambda_1 = \sqrt{2n_c}$  and  $\lambda_2 = \sqrt{n_c + n_u}$ .

Inspired by Pesarin & Salmaso (2010), we also consider a nonparametric combination (NPC) of two dependent permutation tests. Their methodology is based upon properly breaking down a testing problem into a set of simpler subproblems. Then, each subproblem is provided with a proper permutation test, and jointly analyzed to maintain any underlying dependencies. Fitting this approach to our model, we choose a permutation paired  $t$ -test (Janssen, 1999; Konietzschke & Pauly, 2014) that is computed upon the complete pairs  $\mathbf{X}^{(c)}$  only and a permutation Welch-test (Janssen, 1997; Chung et al., 2013; Pauly et al., 2015) that is based upon  $X_{1j}^{(i)}$ , and  $X_{2k}^{(c)}$ . The global  $p$ -value is then obtained through combining the partial  $p$ -values of the above tests using Fisher’s combining function. We denote this testing procedure by  $T_p$ . For more details about the NPC procedure and related R codes, we refer to the monographs of Pesarin (2001) and Pesarin & Salmaso (2010).

Finally, we also consider the most simple solution: the paired  $t$ -test  $T_t$ , calculated on the complete cases  $\mathbf{X}^{(c)}$  only. We compare the finite sample performance of all these methods and the three new bootstrap approaches from Section 4 in the sequel. To judge the performance of all methods, a parametric bootstrap version of the paired  $t$ -test handling full data before introducing missingness has been included in all tables. The corresponding procedure is denoted by  $F$ .

## 6 | SIMULATION STUDY

In this section, we investigate the finite sample behavior of the methods described in Sections 4 and 5 in extensive simulations. All procedures were studied with respect to their

- (i) type-I error rate control at level  $\alpha = 5\%$  and their
- (ii) power to detect deviations from the null hypothesis.

Small- to moderate-sized paired data samples were generated from the model

$$\mathbf{X}_j = \Sigma^{\frac{1}{2}} \boldsymbol{\varepsilon}_j + \boldsymbol{\mu}, \quad j = 1, \dots, n,$$

where  $\boldsymbol{\varepsilon}_j = [\varepsilon_{1j}, \varepsilon_{2j}]^\top$  is an i.i.d. bivariate random vector with mutually independent components and  $E(\boldsymbol{\varepsilon}_1) = \mathbf{0}$  and  $\text{cov}(\boldsymbol{\varepsilon}_1) = I_2$ .

Different choices of symmetric as well as skewed residuals are considered such as standardized normal, exponential, Laplace, and the  $\chi^2$ -distribution with  $df = 30$  degrees of freedom. For the covariance matrix  $\Sigma$ , we considered the choices

$$\Sigma_1 = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \text{ and } \Sigma_2 = \begin{bmatrix} 1 & \sqrt{2}\rho \\ \sqrt{2}\rho & 2 \end{bmatrix}$$

with varying correlation factor  $\rho \in (-1, 1)$ , representing a homoskedastic and a heteroskedastic covariance setting, respectively. The sample sizes were chosen as  $(n_c, n_u) \in \{(10, 10), (30, 10), (10, 30)\}$  under an MCAR mechanism and  $n \in \{10, 20, 30\}$  under an MAR mechanism.

For each scenario, we generated missings as described below: For the *MCAR mechanism*, missing values are inserted randomly to the second component of the bivariate vector  $\mathbf{X}_j$  until a fixed amount of missing values of size  $n_u$  for the second component is achieved.

For the *MAR mechanism*, the probability of being missing on the second component of  $\mathbf{X}_j$  is based on the corresponding value on the first component in the following way: first, we divide  $\mathbf{X}$  into three groups based on their first component values corresponding to a  $2\sigma$ -rule: the first group is given by  $\{\mathbf{X}_j = (X_{1j}, X_{2j}) : X_{1j} \in (-\infty, -2\sigma_1), j = 1, \dots, n\}$ , the second by  $\{\mathbf{X}_j : X_{1j} \in (-2\sigma_1, 2\sigma_1), j = 1, \dots, n\}$ , and the last group by  $\{\mathbf{X}_j : X_{1j} \in (2\sigma_1, \infty), j = 1, \dots, n\}$ , where  $\sigma_1$  is the variance of the first component. Then, we randomly insert missing values on the second component based on the following missing percentages: 15% for group one and three and 30% for the second group.

In order to assess the power of all methods, we set  $\boldsymbol{\mu} = [\delta, 0]^\top$  with shift parameter  $\delta \in \{0, 1/2, 1\}$ . All simulations were operated by means of the statistical computing environment R based on  $n_{sim} = 10,000$  Monte-Carlo runs and  $B = 999$  bootstrap runs (in case of the three bootstrapped methods based upon  $T_W^*$ ,  $T_A^*$ , and  $T_M^*$  and the bootstrapped version of Little's method  $T_L^*$ ). The algorithm for the computation of the  $p$ -value of the parametric bootstrap tests is as follows:

1. For the given incomplete paired data, calculate the observed test statistic, say  $T$ .
2. Estimate the covariance matrix  $\boldsymbol{\Gamma}$  by  $\hat{\boldsymbol{\Gamma}}$ .
3. Generate a bootstrap sample  $\mathbf{X}_j^* = (X_{1j}^*, X_{2j}^*)$  from  $N(\mathbf{0}, \hat{\boldsymbol{\Gamma}})$ ,  $j = 1, \dots, n$ .
4. Insert missing values in an MCAR or MAR manner to the second component of the vector  $\mathbf{X}_j^*$  resulting in  $\mathbf{X}_j^{*(c)}$  and  $\mathbf{X}_k^{*(i)}$  where  $j = 1, \dots, n_c$ ,  $k = 1, \dots, n_u$ .
5. Calculate the value of the test statistic for the bootstrapped sample  $T^*$ .
6. Repeat the Steps 3 and 4 independently  $B = 999$  times and collect the observed test statistic values in  $T_b^*$ ,  $b = 1, \dots, B$ .
7. Finally, estimate the bootstrap  $p$ -value as  $P\text{-value} = \frac{\sum_{b=1}^B I(T_b^* > T)}{B}$ .

Now, the nonparametric bootstrap method that is used for estimating the covariance  $\eta$  of the Fisher's pooling method as suggested by Qi et al. (2019) is as follows:

1. Draw  $n_c$  times with replacement from the pairs  $\mathbf{X}_j^{(c)} = (X_{1j}^{(c)}, X_{2j}^{(c)})$ ,  $j = 1, \dots, n_c$ , and calculate the  $p$ -value  $P_p^*$ .
2. Draw  $n_u$  times with replacement from  $\mathbf{X}_k^{(i)}$ ,  $k = 1, \dots, n_u$ , and calculate the  $p$ -value  $P_{up}^*$ .
3. Replicate Step 1,  $B = 999$  times and collect the observed  $p$ -values of the Wilcoxon signed-rank test (paired data) and Mann-Whitney U test (unpaired data) in  $P_{pb}^*$  and  $P_{ub}^*$ , respectively,  $b = 1, \dots, B$ .
4. Finally, estimate the parameter  $\eta$  needed for estimating the degrees of freedom as  $\eta = \widehat{\text{cov}}(-2\log(\mathbf{P}_p^*), -2\log(\mathbf{P}_{up}^*))$ , where  $\mathbf{P}_p^* = \{P_{pb}^*, b = 1, \dots, B\}$  and  $\mathbf{P}_{up}^* = \{P_{ub}^*, b = 1, \dots, B\}$ .

**Type-I Error Results.** Simulation results of type-I error level of the studied procedures under the MCAR framework for different sample sizes and for homoskedastic as well as heteroskedastic settings are summarized in Tables 1, S.1, and S.2.

It can be readily seen that the suggested bootstrap approaches based upon  $T_W^*$ ,  $T_A^*$  and  $T_M^*$  tend to result in quite accurate type-I error rate control under homoskedasticity as well as heteroskedasticity and over the whole range of correlation factors for most settings. Only in two cases; First, in case of the negative unbalanced sample size (10,30), particularly under heteroskedasticity, the bootstrapped MATS ( $T_M^*$ ) is not recommended due to its liberal behavior. However, in this case, the other two suggested bootstrapped tests  $T_W^*$ , and  $T_A^*$  are controlling type-I error rate accurately. Secondly, in case of the skewed exponential distribution, the control is not adequate and a liberal behavior is observed. However, in this case, all the other chosen procedures also failed to control type-I error rate for the underlying sample sizes, which are indicated in bold red through all tables. Specifically, in the case of homoskedasticity, and a balanced sample size (10,10), our three suggested tests still result in accurate test decisions. For a positive balanced sample size (30,10), the bootstrapped ATS ( $T_A^*$ ) still controls type-I error rate accurately under homoskedastic as well heteroskedastic settings. It has even the best control of type-I error rate under heteroskedasticity among all considered methods that are identified by bold entries in the table.

TABLE 1 Type-I error simulation results ( $\alpha = .05$ ) of the tests for different distributions under varying correlation values ( $\rho$ ) with sample sizes  $(n_c, n_u) = (10, 10)$  and different covariance matrices  $\Sigma_1$  and  $\Sigma_2$  under the MCAR framework

Dist	$\rho$	$\Sigma_1$									$\Sigma_2$								
		F	Parametric bootstrap				Alternatives				F	Parametric bootstrap				Alternatives			
			$T_W^*$	$T_A^*$	$T_M^*$	$T_L^*$	$T_t$	$T_N$	$T_F$	$T_P$		$T_W^*$	$T_A^*$	$T_M^*$	$T_L^*$	$T_t$	$T_N$	$T_F$	$T_P$
Normal	-.9	5.3	5.3	5.2	5.4	<b>5.0</b>	4.8	6.7	4.3	<b>7.5</b>	5.1	<b>5.0</b>	5.3	5.6	4.9	4.7	<b>7.1</b>	4.8	<b>8.3</b>
	-.5	5.3	5.3	5.7	5.6	5.3	<b>5.1</b>	<b>6.8</b>	4.6	<b>7.4</b>	5.3	5.3	5.5	6	5.2	<b>5.0</b>	<b>6.8</b>	5.1	<b>8</b>
	-.1	5.3	4.6	4.9	4.8	4.8	<b>5.0</b>	6.5	4	6.3	5.4	4.6	5.5	5.2	<b>5.0</b>	5.1	<b>7.2</b>	4.6	<b>7.7</b>
	.1	4.9	4.8	5.4	5.1	<b>5.0</b>	4.8	6.4	4.3	6.6	<b>4.9</b>	<b>5.1</b>	5.5	5.4	<b>5.1</b>	<b>4.9</b>	<b>7</b>	<b>4.9</b>	<b>7.7</b>
	.5	5.3	5.4	<b>5.1</b>	<b>5.1</b>	4.3	5.3	6.2	4.3	5.9	5.2	5.4	5.6	5.5	<b>4.9</b>	5.2	<b>7.1</b>	4.7	<b>7.3</b>
	.9	5.3	5.2	<b>5.0</b>	4.3	4.5	5.4	5.7	4.2	<b>5.0</b>	5.2	5.1	<b>5.0</b>	5.9	3.2	5.4	<b>6.9</b>	4.5	<b>7.1</b>
Laplace	-.9	<b>4.9</b>	4.4	<b>4.9</b>	5.5	4.6	4.8	6.5	4.5	<b>7.5</b>	<b>4.9</b>	4.4	<b>5.1</b>	5.7	4.4	4.7	<b>7.1</b>	4.8	<b>8</b>
	-.5	5.1	4.4	5.2	5.1	<b>5.0</b>	4.5	6.6	4.3	<b>7.3</b>	5.1	4.4	5.1	5.4	<b>5.0</b>	4.6	<b>7</b>	4.7	<b>7.9</b>
	-.1	<b>4.9</b>	4.2	<b>4.9</b>	4.8	4.6	4.6	6.4	4.2	6.5	<b>5.0</b>	4.4	<b>5.0</b>	<b>5.0</b>	4.6	4.5	<b>6.8</b>	4.5	<b>7.4</b>
	.1	<b>4.8</b>	4.3	4.3	4.2	4.3	4.4	6.2	4	6.1	<b>4.9</b>	4.3	4.6	4.6	4.3	4.5	6.6	4.3	<b>7</b>
	.5	<b>5.1</b>	4.4	4.5	4.4	3.6	4.5	6.2	4	5.8	<b>4.9</b>	4.4	4.6	4.5	3.7	4.5	6.7	4.2	<b>6.8</b>
	.9	4.8	3.9	4.8	3.6	4.7	4.4	5.6	4	<b>4.9</b>	<b>4.8</b>	4.1	4.7	5.4	3.9	4.4	6.6	4.1	6.4
Exponential	-.9	<b>4.8</b>	4.7	4.4	5.6	4.5	4.2	6.5	4.3	<b>6.8</b>	<b>5.0</b>	4.6	5.2	<b>6.8</b>	5.3	4.7	<b>8.7</b>	4.1	<b>8.2</b>
	-.5	5.2	<b>5.1</b>	<b>4.9</b>	4.8	5.3	4.2	6.4	4.4	<b>7.1</b>	5.4	5.4	6.4	6.3	6.7	<b>5.0</b>	<b>9.7</b>	4.8	<b>8.9</b>
	-.1	5.3	5.3	<b>5.0</b>	4.6	5.8	4.4	6.6	4.3	<b>6.9</b>	5.6	6.1	6.6	6.1	<b>7.1</b>	<b>5.1</b>	<b>10.1</b>	4.8	<b>8.7</b>
	.1	<b>5.0</b>	<b>5.0</b>	4.4	4.1	5.9	4	6.6	4.1	6.2	5.6	6.1	<b>6.8</b>	5.9	<b>7.6</b>	5.2	<b>10.5</b>	<b>5.0</b>	<b>8.7</b>
	.5	<b>5.1</b>	5.8	4.5	4.2	6.5	4.2	6.2	4.7	6.1	5.9	<b>7</b>	<b>6.9</b>	6.5	<b>7.7</b>	5.7	<b>10.7</b>	<b>5.1</b>	<b>8.8</b>
	.9	<b>4.7</b>	5.8	4.4	3.6	<b>7.3</b>	4.1	5.5	4.4	4.4	<b>7.5</b>	<b>8</b>	<b>5.4</b>	<b>8.5</b>	<b>7.8</b>	<b>8.8</b>	<b>12.1</b>	5.6	<b>9.7</b>
Chi-square	-.9	<b>5.2</b>	5.4	5.6	5.8	<b>5.2</b>	<b>5.2</b>	<b>6.9</b>	<b>4.8</b>	<b>7.8</b>	5.4	5.5	5.8	6.1	<b>5.0</b>	5.3	<b>7.6</b>	5.2	<b>8.7</b>
	-.5	5.4	<b>5.0</b>	5.2	5.1	<b>5.0</b>	4.9	6.5	4.1	<b>6.9</b>	5.4	<b>5.0</b>	5.3	5.6	<b>5.0</b>	4.8	<b>7.3</b>	4.5	<b>7.6</b>
	-.1	5.1	<b>5.0</b>	5.1	5.3	5.3	4.9	6.4	4.4	6.6	<b>5.0</b>	5.1	5.6	5.7	5.4	4.9	<b>7</b>	4.5	<b>7.9</b>
	.1	5.3	<b>5.0</b>	5.1	5.1	<b>5.0</b>	5.1	6.6	4.4	6.4	5.4	<b>5.0</b>	5.6	5.7	5.2	5.1	<b>6.9</b>	4.4	<b>7.6</b>
	.5	5.3	5.4	<b>5.0</b>	<b>5.0</b>	4.4	5.1	6.7	4.3	5.9	5.3	5.4	5.3	5.3	4.6	<b>5.0</b>	6.6	4.5	<b>7</b>
	.9	5.2	<b>5.0</b>	5.3	4.1	4.6	4.8	5.9	4.3	5.1	<b>5.2</b>	5.3	5.6	6.5	3.3	5.4	<b>7.6</b>	4.3	<b>7.6</b>

Note. For each setting, the values closest to the prescribed level are printed in bold and values exceeding the upper limit (6.8%) of the 99% binomial interval are in red color.

Moreover, the bootstrapped test that is based on the maximum likelihood estimator  $T_L^*$  tends to behave similar to our three suggested bootstrap procedures in controlling type-I error rate. Only in the case of large positive correlation factors  $\rho = .9$ , it results in very conservative decisions.

In contrast, the other tests ( $T_N, T_F, T_P$ ) do not control type-I error level constantly under heteroskedasticity or even under homoskedasticity in all of the considered sample sizes. It can also be seen from Tables 1, S.1, and S.2 that the NCT  $T_N$ , controls type-I error quite accurately in the case of larger numbers of complete pairs ( $n_c = 30$ ), but it shows liberal behavior for smaller numbers of complete pairs ( $n_c = 10$ ). This test turns very liberal in the case of heteroskedasticity. Furthermore, the FPM test  $T_F$  tends to result in a quite accurate type-I error control in the case of smaller numbers of complete pairs. For larger numbers of complete pairs, it leads to a conservative decision. For these scenarios, this behavior does not depend on the homoskedasticity assumption. Moreover, the NPC  $T_P$  shows a quite liberal behavior in most of the considered settings. Regarding the paired  $t$ -test based on the complete observations  $T_t$ , an inflation of the type-I error rate could be realized for certain distributions, when the missing rate was large and the number of complete pairs was small, see for example, the scenario  $(n_c, n_u) = (10, 10)$ . The effect vanishes for a larger number of complete pairs. This is in line with the theoretical results of the paired  $t$ -test with i.i.d. observations. The results also indicate that the paired bootstrapped  $t$ -test on the full data  $F$  controls type-I error through almost all settings.

It was also interesting to discover the type-I error rate control of the tests under similar attributes to the breast cancer gene study data which reflects data sets with a few pairs and large amount of unpaired portions. Simulation results for the type-I error rate of the studied procedures for  $(n_c = 16, n_u = 74)$  sample sizes are presented in Tables S.22 and S.23. The correlation  $\rho$  in Table S.23 is estimated based on the data. It can be easily seen from Tables S.22 and S.23 that the bootstrap tests are

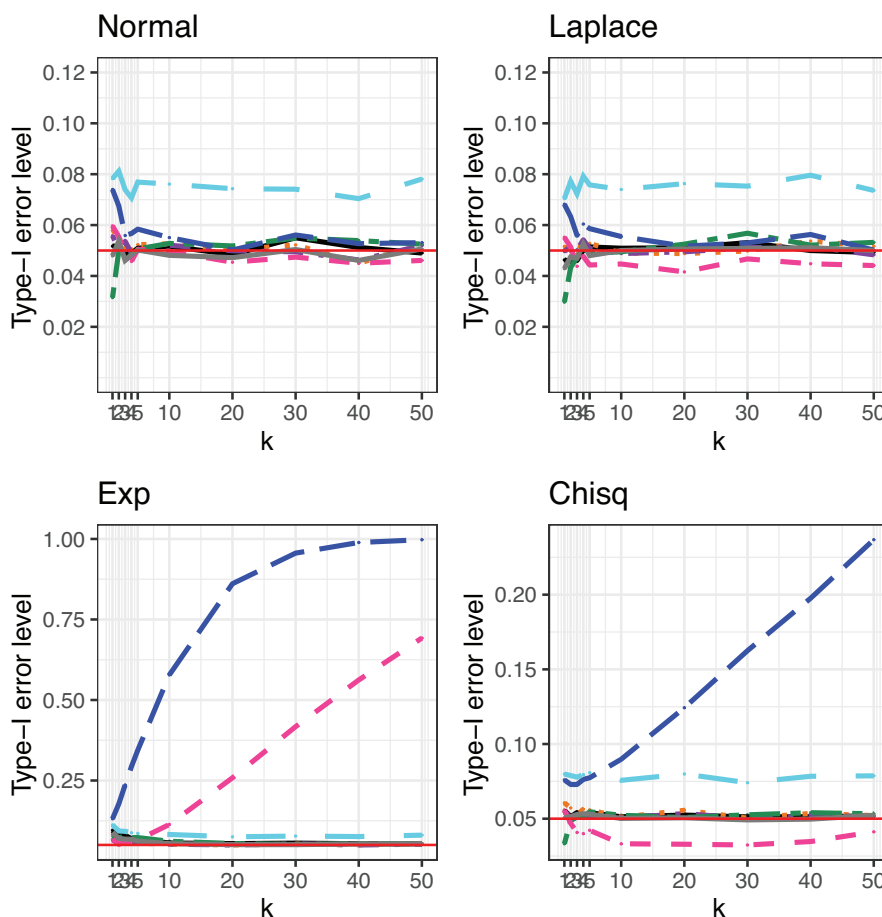


FIGURE 2 Type-I error simulation results ( $\alpha = .05$ ) of the tests  $T_W^*$  (—),  $T_A^*$  (· · ·),  $T_M^*$  (— · —),  $T_L^*$  (— · —),  $T_N$  (— —),  $T_F$  (— · —),  $T_I$  (— —), and  $T_P$  (— · —) for different distributions under correlation factor ( $\rho = .9$ ) and heteroskedastic covariance matrix  $\Sigma_2$  for varying  $k$  values multiplied by  $(n_c, n_u) = (10, 30)$  under the MCAR framework

robust under large amounts of missing observations and control type-I error rate accurately, especially the bootstrapped tests  $T_W^*$ , and  $T_A^*$ . Except in the case of exponential distribution. The alternative approach  $T_N$  has acceptable control under homoskedasticity. But, under the exponential distribution, it turned very liberal especially under heteroskedasticity, while the Fisher’s pooling method tends to result in quite acceptable control in most cases.

Simulation results of the type-I error level of the studied procedures under the MAR framework for different sample sizes and covariance structures are summarized in Tables S.3– S.5. There, it can be seen that for moderate to large sample sizes ( $n \in \{20, 30\}$ ), the bootstrapped ATS  $T_A^*$ , the bootstrapped WTS  $T_W^*$ , the bootstrapped MATS  $T_M^*$ , the bootstrapped Little  $T_L^*$ , and the NCT  $T_N$  exhibit a fairly good type-I error rate control for almost all considered scenarios under homoskedasticity as well as heteroskedasticity. Only in the case of the skewed exponential distribution, the control of  $T_W^*$ ,  $T_M^*$ , and  $T_N$  is not adequate and liberal behavior is observed, which is marked with red through all tables. In contrast, the bootstrapped MATS  $T_M^*$  tends to be sensitive to the dependency structure in the data. In particular,  $T_M^*$  exhibits a liberal behavior for negative correlations. For small sample sizes ( $n = 10$ ), the  $T_N$  test tends to be liberal in all considered situations. In contrast, the bootstrapped tests  $T_W^*$ ,  $T_M^*$ , and  $T_L^*$  exhibit good type-I error rate control for most settings except for the Laplace distribution. The bootstrapped ATS  $T_A^*$  tends to be very conservative especially under heteroskedasticity. However, the FPM  $T_F$  exhibits a conservative behavior under most considered situations.

**Further Investigations on Type-I Error.** In addition to the small and moderate sample size settings, we were also interested in studying type-I error rate control when *sample sizes increase*, while missing rates remain nearly unchanged. For moderate to large sample sizes, we considered the choices  $(n_c, n_u) = k \cdot (10, 30)$  and  $(n_c, n_u) = k \cdot (1, 1) + (10, 10)$ , where  $k$  ranges from 1 to 50 (balanced case) and 0 to 500 (unbalanced case), respectively. Figures 2 and S.1 summarize the type-I error rate ( $\alpha = .05$ ) for these settings. The results indicate that the NCT by Qi et al. (2019)  $T_N$  controls type-I error rate quite accurate under symmetric distributions, however, it fails to control type-I error rate under skewed distributions.

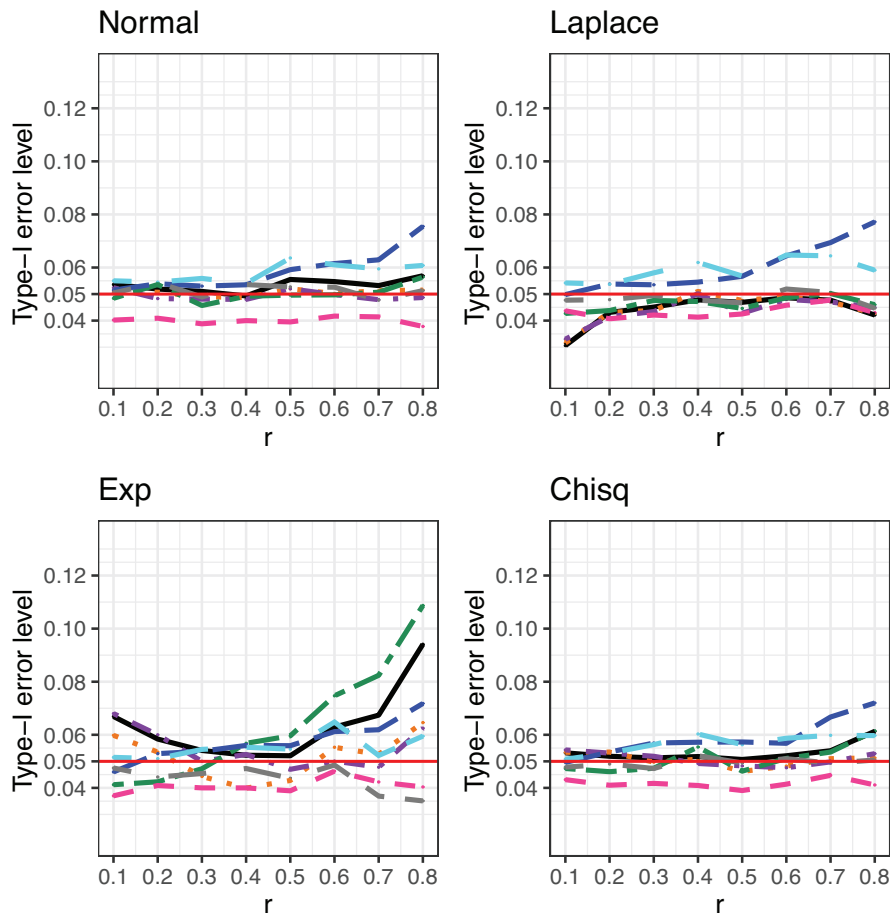
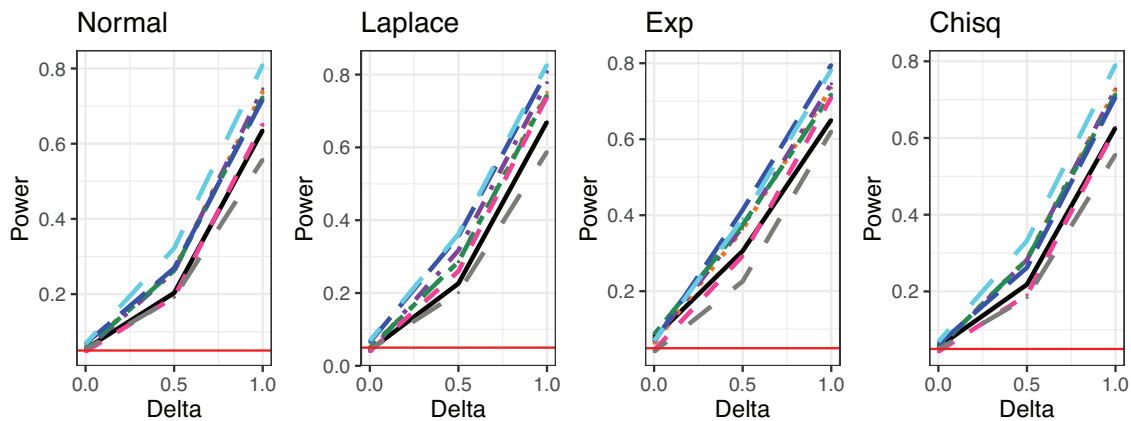


FIGURE 3 Type-I error simulation results ( $\alpha = .05$ ) of the tests  $T_W^*$  (—),  $T_A^*$  (· · ·),  $T_M^*$  (— · —),  $T_L^*$  (— · —),  $T_N$  (— —),  $T_F$  (— · —),  $T_t$  (— —), and  $T_P$  (— · —) for different distributions under correlation factor ( $\rho = .5$ ) with sample size ( $n = 30$ ) and homoskedastic covariance matrix  $\Sigma_1$  for varying missing rates  $r \in \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8\}$  under the MCAR framework

In fact, it gets even more liberal with increasing sample sizes. In contrast, the FPM  $T_F$  by Qi et al. (2019) tends to be conservative when missing rates among subjects of 50% are present, even under large numbers of complete observations such as  $n_c = 510$  (Figure S.1). For larger missing rates (75%), it shows surprisingly quite accurate type-I error control (see Figure 2). Only in case of the exponential distribution, a very liberal behavior is observed that is acting analogous to a power function with increment of sample sizes (Figure 1). Here, the suggested bootstrap approaches  $T_A^*$ ,  $T_W^*$ ,  $T_M^*$ , and  $T_L^*$  are the only methods that control type-I error rate accurately among all considered settings. The  $t$ -test  $T_t$  based on the complete cases controls type-I error as well, but had challenges with small complete cases  $n_c \leq 10$ . The NPC-test  $T_P$ , however, revealed a constant inflation of the type-I error rate for all missing rate scenarios. The degree of inflation remained the same even for increasing missing rates. Therefore,  $T_P$  seems not to be an adequate choice, even for smaller missing rates.

In order to cover the effect of increasing missing rates, we studied type-I error control for sample sizes of the form  $(n_c, n_u) = ((1 - r) \cdot 30, r \cdot 30)$  with  $r \in \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8\}$  covering missing rates (among subjects) from 10% to 80% under moderate positive correlation factor ( $\rho = .5$ ). Figures 3 and S.2 summarize type-I error rate control for these settings under a homoskedastic and a heteroskedastic covariance structure, respectively. The results indicate that under homoskedasticity, the alternative approach  $T_N$  tends to be slightly liberal. It moves closer to the 0.05 threshold for missing rates below 60%. In contrast, under heteroskedasticity,  $T_N$  tends to be more sensitive to the missing rates. In particular, it exhibits a conservative or liberal behavior for lower and larger missing rates, respectively. However, under this moderate sample size ( $n = 30$ ) and correlation factor ( $\rho = .5$ ), the FPM  $T_F$  tends to be conservative under all considered settings and its behavior is independent of the missing rate or even homoskedasticity assumption. In contrast, the suggested bootstrap approaches tend to control type-I error rate more accurate over the range of missing rates  $r$  for most settings. Only in case





**FIGURE 4** Power simulation results ( $\alpha = .05$ ) of the tests  $T_W^*$  (—),  $T_A^*$  (···),  $T_M^*$  (- · -),  $T_L^*$  (- - -),  $T_N$  (— —),  $T_F$  (- - -),  $T_t$  (— —), and  $T_P$  (- - -) for different distributions under correlation factor ( $\rho = .1$ ) with sample size  $(n_c, n_u) = (10, 30)$  and homoskedastic covariance matrix  $\Sigma_1$  under the MCAR framework

of the skewed exponential distribution and missing rates greater than 50%, the control is not adequate. However, in this case all the other chosen procedures also failed to control the type-I error rate.

**Power.** In addition to the type-I error rate control, we studied the power of the nine tests for all considered settings. Figure 4 summarizes the power simulation results for a negative balanced sample size (10,30) under the MCAR framework. The power simulation results for the other scenarios are included in the supplement. The power analysis results of the considered methods under MCAR and MAR frameworks involving homoskedastic as well as heteroskedastic settings are summarized in Tables S6–S11 in supplement for the MCAR mechanism and Tables S12–S17 in supplement for the MAR mechanism. The entries that belong to very liberal tests have been colored in red in the power tables. It can be readily seen that the four bootstrapped tests  $T_W^*$ ,  $T_A^*$ ,  $T_M^*$ , and  $T_L^*$  and the NCT  $T_N$  have almost similar large power behavior under homoskedastic as well as heteroskedastic settings. Only in the heteroskedastic cases with skewed exponential distribution, the NCT  $T_N$  shows larger power than the others, which is due to its rather liberal behavior. One should also notice that the power behavior of each test varies based on the dependency structure of the data except for the bootstrapped ATS  $T_A^*$ . As expected, the paired  $t$ -test based on complete observations  $T_t$  revealed for small complete observations low power results compared to the other approaches. The NPC-test  $T_P$  also shows larger power results, but the effect can be led back to its liberal type-I error behavior.

## 7 | ILLUSTRATIVE DATA ANALYSES

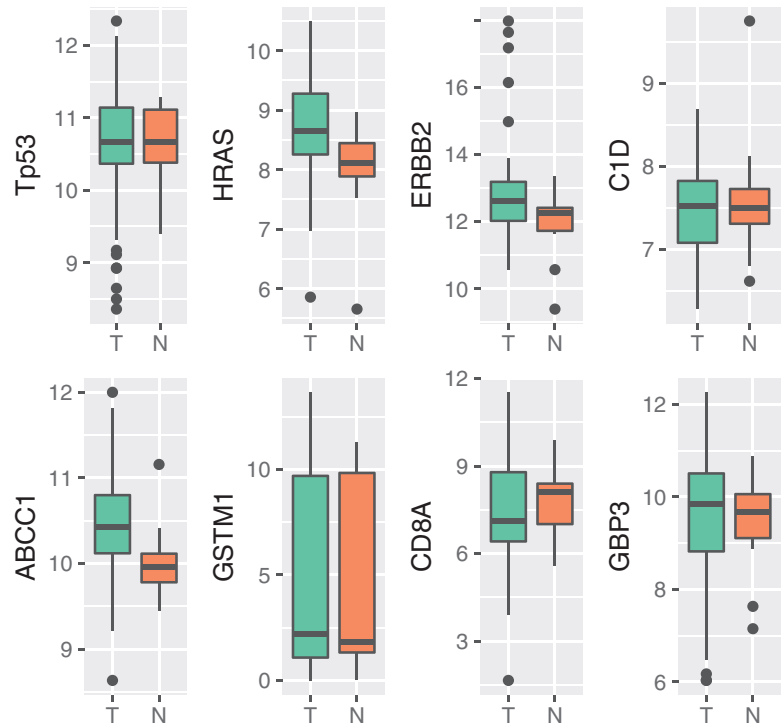
In this section, we consider three real life problems coming from different sectors and sources. We start with a genome study on breast cancer.

### 7.1 | Breast cancer study: gene expression data

The TCGA project is a pilot project which was launched in 2005 with a financial support from the National Institutes of Health. It aims to understand the genetic basis of several types of human cancers through the application of high-throughput genome analysis techniques. TCGA collects molecular information such as miRNA/mRNA expressions, protein expressions, and weight of the sample as well as clinical data about the patients.

A breast cancer study has been performed by TCGA to improve the ability of diagnosing, treating, and preventing breast cancer through investigating the genetic basis of carcinoma. Their study consists of 1093 breast cancer patients with Clinical and RNA sequencing records. Among them, there were 112 subjects that provided both, normal, and tumor tissues. Here, we were interested in a subset of this datum that contains patients with pathologic stage I. This subset contains a total of  $n_c = 16$  complete pairs and an unpaired sample for the patients who developed only tumor tissues of size  $n_u = 74$ . The data can be downloaded from Firehouse ([www.gdac.broadinstitute.org](http://www.gdac.broadinstitute.org)).

**FIGURE 5** Profile of the gene expression levels of the tumor and normal breast tissues



**TABLE 2** Unadjusted two-sided  $p$ -values of the breast cancer study

Gene	Parametric bootstrap				Alternatives			
	$T_W^*$	$T_A^*$	$T_M^*$	$T_L^*$	$T_t$	$T_N$	$T_F$	$T_P$
TP53	0.928	0.852	0.903	0.877	0.689	0.954	0.949	0.901
ABCC1	0.002	0.003	0.002	0.002	0.365	0.003	0.004	0
HRAS	0.007	0.002	0.003	0.002	0.022	0.001	0.004	0
GSTM1	0.821	0.85	0.849	0.515	0.605	0.629	0.967	0.827
ERBB2	0.043	0.024	0.011	0.014	0.136	0.071	0.069	0.007
CD8A	0.463	0.51	0.484	0.434	0.885	0.555	0.468	0.53
C1D	0.772	0.553	0.622	0.555	0.553	0.587	0.792	0.608
GBP3	0.196	0.301	0.214	0.084	0.083	0.103	0.357	0.262

Based on previous studies, six genes have been found to be significantly associated with breast cancer: **TP53**, **ABCC1**, **HRAS**, **GSTM1**, **ERBB2**, and **CD8A** (Harari & Yarden, 2000; De Jong et al., 2002; Munoz et al., 2007; Finak et al., 2008). Another two genes: **C1D** and **GBP3** were under investigation although they did not show any significant relation toward breast cancer patients (Qi et al., 2019). In this paper, we aim to test the hypothesis whether mean genetic expressions of the eight genes are significantly different between normal and tumor tissues for patients with early stage I breast cancer. Boxplots representing the characteristics of the eight genes are shown in Figure 5.

We applied all bootstrap testing methods  $T_W^*$ ,  $T_A^*$ ,  $T_M^*$ , and  $T_L^*$  as well as the alternative approaches  $T_t$ ,  $T_N$ ,  $T_F$ , and  $T_P$  to detect the null hypothesis of equal means between normal and tumor tissues ( $H_0 : \mu_1 = \mu_2$ ) against the two-sided alternative ( $H_1 : \mu_1 \neq \mu_2$ ). The results are summarized in Table 2.

It can be seen from Table 2 that the bootstrapped approaches  $T_W^*$ ,  $T_A^*$ ,  $T_M^*$ , and  $T_L^*$  and the NPC  $T_P$  identified three of eight genes having significantly different genetic expressions in normal and tumor tissues; genes **ABCC1**, **HRAS**, and **ERBB2**. However, the NCT  $T_N$ , and the FPM  $T_F$  led to different results for the **ERBB2** gene. Regarding the paired  $t$ -test based on the complete observations  $T_t$ , different results obtained for the **ABCC1** and **ERBB2** genes.



TABLE 3 Two-sided  $p$ -values of the considered studies

Study	$F$	Parametric bootstrap				Alternatives			
		$T_W^*$	$T_A^*$	$T_M^*$	$T_L^*$	$T_t$	$T_N$	$T_F$	$T_P$
Anorexia	0.002	0.026	0.043	0.029	0.03	0.004	0.136	0.008	0.022
GrapeFruit	0.002	0.039	0.014	0.029	0.068	0.141	0.031	0.068	0.022

## 7.2 | Two more examples

To illustrate potential differences between all methods we consider two additional examples called “Anorexia” and “GrapeFruit.” Each of them consists of complete data sets and missing values were introduced on them by the MCAR mechanism with a missing rate of  $r = 30\%$ . They can be briefly described as follows:

**Anorexia.** This data set consists of weights in pounds for 17 young girls who were receiving a treatment for anorexia over a fixed period of time. The main problem is to compare the girls’ weights before and after the treatment. This datum was originally published by Hand et al. (1993), and were analyzed in Pruzek & Helmreich (2009). It is also included in the R package PairedData (Champely & Champely, 2018).

**GrapeFruit.** It consists of a paired samples data that are taken from Preece (1982). The study aimed to detect differences between “shaded” and “exposed” grapefruits. To make the differences as precise as possible, they dealt with both halves of a single fruit under similar conditions. This data set consists of the percentages of solids in the shaded and exposed halves of 25 grapefruits. This datum is also contained in the R package PairedData (Champely & Champely, 2018).

We applied the  $F$ -test that considers the full data before missingness, all bootstrapped approaches  $T_W^*$ ,  $T_A^*$ ,  $T_M^*$ , and  $T_L^*$  as well as the alternative approaches  $T_t$ ,  $T_N$ ,  $T_F$ , and  $T_P$  to detect the null hypothesis of equal means  $H_0 : \{\mu_1 = \mu_2\}$  against the two-sided alternative  $H_1 : \{\mu_1 \neq \mu_2\}$ . The results are summarized in Table 3. It can be seen from Table 3 that the full data test  $F$ , bootstrapped approaches  $T_W^*$ ,  $T_A^*$ , and  $T_M^*$ , and the Pesarin test  $T_P$  identified significant differences in both data sets. However, the alternative naive approach based on the complete observations  $T_t$  and the FPM  $T_F$  failed in detecting significant difference in the GrapeFruit data set. In addition, the NCT  $T_N$  could not identify any significant difference for the Anorexia data set.

## 8 | DISCUSSION AND OUTLOOK

The problem of matched pairs with missing values occurs frequently in practice. Most available procedures in the literature are not applicable when missing values occur in a single arm. Exceptions are given by the recent NCT and FPM approaches of Qi et al. (2019). For the NCT approach, Qi et al. (2019) utilize a combination of the sign and Wilcoxon Mann–Whitney rank sum test. And, the FPM approach, is based on a weighted combination of the  $p$ -values of the Wilcoxon signed rank test and the Wilcoxon Mann–Whitney rank sum test. For homoskedastic settings with symmetric distributions, the NCT and FPM approaches can be recommended. If, however, the underlying assumptions are not true (e.g., in skewed heteroskedastic setups), the NCT and FPM may result in highly inflated type-I errors or considerable power loss. In addition to the NCT and FPM approaches, we also studied a single-arm missingness modification of a nonparametric testing procedure given in Pesarin & Salmaso (2010). It is based on the usage of the permutation paired  $t$ -test and the permutation Welch test on partial combination of the whole data  $D_n$  with missingness. However, the proposed combination strategy did not reveal favorable results leading to a constant inflation of the type-I error. We also calculated the paired  $t$ -test based on complete observations only.

To overcome all these issues, we have provided resampling procedures that are not based on any parametric assumptions and use all observed information within the matched pairs design. They were shown to be asymptotically correct and robust under heteroskedasticity and skewed distributions. The tests were based on restructuring all observed information in a test statistic of quadratic form that can be either a WTS, an ATS, or a MATS. Since WTS is well known (from other situations like in Vallejo et al., 2010; Konietschke et al., 2015; or Smaga, 2017) for being liberal, while ATS and MATS tend to be rather conservative or liberal for small to moderate sample sizes, we improved their small sample behavior by an asymptotic model-based bootstrap approach. The procedure’s asymptotic validity was also proven and can be found in the supplement. In addition, we improved the behavior of the Little’s test (cf. Little, 1976) that is based upon a modified maximum likelihood estimator by introducing an asymptotic model-based bootstrap version of the test.

In an extensive simulation study, the type-I error rate control of the tests have been examined for symmetric and skewed distributions with homoskedastic and heteroskedastic covariance settings under different missing mechanisms. There, it was seen that the parametric bootstrap versions of WTS, ATS, MATS, and Little improve their small sample behavior. In particular, our bootstrap tests have been shown to perform very well in most of the cases, even with larger amount of missingness, heteroskedastic covariance or skewed data. Only the type-I error control for the exponential distribution, particularly under heteroskedasticity, MCAR and small paired sample sizes with rather large unpaired portions ( $n_c = 10, n_u = 30$ ), is not maintained. In this setting, however, all other considered methods such as the ones given in Qi et al. (2019) and inspired by Pesarin (2001) and Pesarin & Salmaso (2010) also failed to control the type-I error rate.

Furthermore, our simulation study reveals that the bootstrap procedures' type-I error control is not much affected by less stringent missing data mechanism such as the MAR. However, their power behavior is affected. A possible justification of the latter effect might originate from the additional dependence structure within the occurrence of missing values compared to the MCAR case. It seems that the testing procedure is more challenged to detect deviations from the null.

In order to simplify the application of our approaches, the three proposed bootstrap statistical methods have been implemented within the PBT function in the freely available R-package **MissPair**. It is available on GitHub (<https://github.com/lubnaamro/MissPair>) and will be available on the CRAN repository.

Future research will be concerned with extending our procedures to multivariate settings (MANOVA). An investigation of the behavior of our methods together with *logit* or *probit* transformations may also be part of future work.

## ACKNOWLEDGMENTS

Burim Ramosaj and Markus Pauly acknowledge the support of the German Research Foundation (DFG). Lubna Amro's work was also supported by the German Academic Exchange Service (DAAD) under the project: Research Grants - Doctoral Programmes in Germany, 2015/16 (No. 57129429).


## CONFLICT OF INTEREST

The authors have declared no conflict of interest.

## DATA AVAILABILITY STATEMENT

The data that supports the findings of this study are available in the supplementary material of this article.

## OPEN RESEARCH BADGES

 This article has earned an Open Data badge for making publicly available the digitally-shareable data necessary to reproduce the reported results. The data is available in the [Supporting Information](#) section.

This article has earned an open data badge “**Reproducible Research**” for making publicly available the code necessary to reproduce the reported results. The results reported in this article could fully be reproduced.

## ORCID

Lubna Amro  <https://orcid.org/0000-0001-8550-7406>

Markus Pauly  <https://orcid.org/0000-0002-0976-7190>

Burim Ramosaj  <https://orcid.org/0000-0002-1885-5168>

## REFERENCES

- Amro, L., Konietschke, F., & Pauly, M. (2019). Multiplication-combination tests for incomplete paired data. *Statistics in Medicine*, *38*, 3243–3255.
- Amro, L., & Pauly, M. (2017). Permuting incomplete paired data: A novel exact and asymptotic correct randomization test. *Journal of Statistical Computation and Simulation*, *87*, 1148–1159.
- Bhoj, D. S. (1978). Testing equality of means of correlated variates with missing observations on both responses. *Biometrika*, *65*, 225–228.
- Brown, M. B. (1975). 400: A method for combining non-independent, one-sided tests of significance. *Biometrics*, *31*, 987–992.
- Brunner, E. (2001). Asymptotic and approximate analysis of repeated measures designs under heteroscedasticity. *Mathematical Statistics with Applications in Biometry, Festschrift in Honour of Prof. Dr. Siegfried Schach*, 313–326.
- Champely, S., & Champely, M. S. (2018). R-Package “PairedData” Version 1.1.1.
- Chung, E., & Romano, J. P. (2013). Exact and asymptotically robust permutation tests. *Annals of Statistics*, *41*, 484–507.
- De Jong, M., Nolte, I., Te Meerman, G., Van der Graaf, W., Oosterwijk, J., Kleibeuker, J., Schaapveld, M., & De Vries, E. (2002). Genes other than *BRCA1* and *BRCA2* involved in breast cancer susceptibility. *Journal of Medical Genetics*, *39*, 225–242.
- Ekbohm, G. (1976). On comparing means in the paired case with incomplete data on both responses. *Biometrika*, *63*, 299–304.

- Feng, Q., Hawes, S. E., Stern, J. E., Wiens, L., Lu, H., Dong, Z. M., Jordan, C. D., Kiviati, N. B., & Vesselle, H. (2008). DNA methylation in tumor and matched normal tissues from non-small cell lung cancer patients. *Cancer Epidemiology and Prevention Biomarkers*, *17*, 645–654.
- Finak, G., Bertos, N., Pepin, F., Sadekova, S., Souleimanova, M., Zhao, H., Chen, H., Omeroglu, G., Meterissian, S., Omeroglu, A., Hallett, M., & Park, M. (2008). Stromal gene expression predicts clinical outcome in breast cancer. *Nature Medicine*, *14*, 518–527.
- Friedrich, S., & Pauly, M. (2018). Mats: Inference for potentially singular and heteroscedastic MANOVA. *Journal of Multivariate Analysis*, *165*, 166–179.
- Fuchs, N., Pölz, W., & Bathke, A. C. (2017). Confidence intervals for population means of partially paired observations. *Statistical Papers*, *58*, 35–51.
- Hand, D. J., Daly, F., McConway, K., Lunn, D., & Ostrowski, E. (1993). *A handbook of small data sets*. CRC Press.
- Harari, D., & Yarden, Y. (2000). Molecular mechanisms underlying ErbB2/HER2 action in breast cancer. *Oncogene*, *19*, 6102–6114.
- Harrar, S. W., Feyasa, M. B., & Wencheke, E. (2020). Nonparametric procedures for partially paired data in two groups. *Computational Statistics & Data Analysis*, *144*, 106903.
- Hartung, J. (1999). A note on combining dependent tests of significance. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, *41*, 849–855.
- Hou, C.-D. (2005). A simple approximation for the distribution of the weighted combination of non-independent or independent probabilities. *Statistics & Probability Letters*, *73*, 179–187.
- Janssen, A. (1997). Studentized permutation tests for non-i.i.d. hypotheses and the generalized Behrens-Fisher problem. *Statistics & Probability Letters*, *36*, 9–21.
- Janssen, A. (1999). Nonparametric symmetry tests for statistical functionals. *Mathematical Methods of Statistics*, *8*, 320–343.
- Kim, B. S., Kim, I., Lee, S., Kim, S., Rha, S. Y., & Chung, H. C. (2004). Statistical methods of translating microarray data into clinically relevant diagnostic information in colorectal cancer. *Bioinformatics*, *21*, 517–528.
- Konietschke, F., Bathke, A. C., Harrar, S. W., & Pauly, M. (2015). Parametric and nonparametric bootstrap methods for general MANOVA. *Journal of Multivariate Analysis*, *140*, 291–301.
- Konietschke, F., & Pauly, M. (2014). Bootstrapping and permuting paired t-test type statistics. *Statistics and Computing*, *24*, 283–296.
- Kost, J. T., & McDermott, M. P. (2002). Combining dependent p-values. *Statistics & Probability Letters*, *60*, 183–190.
- Krishnamoorthy, K., & Lu, F. (2010). A parametric bootstrap solution to the MANOVA under heteroscedasticity. *Journal of Statistical Computation and Simulation*, *80*, 873–887.
- Kuan, P. F., & Huang, B. (2013). A simple and robust method for partially matched samples using the p-values pooling approach. *Statistics in Medicine*, *32*, 3247–3259.
- Kuriakose, M. A., Chen, W. T., He, Z. M., Sikora, A. G., Zhang, P., Zhang, Z. Y., Qiu, W. L., Hsu, D. F., McMunn-Coffran, C., Brown, S. M., Elango, E. M., Delacure, M. D., & Chen, F. A., et al. (2004). Selection and validation of differentially expressed genes in head and neck cancer. *Cellular and Molecular Life Sciences: CMLS*, *61*, 1372–1383.
- Lapointe, J., Li, C., Higgins, J. P., Van De Rijn, M., Bair, E., Montgomery, K., Ferrari, M., Egevad, L., Rayford, W., Bergerheim, U., Ekman, P., DeMarzo, A., Tibshirani, R., Botstein, D., Brown, P., Brooks, J., & Pollack, J. (2004). Gene expression profiling identifies clinically relevant subtypes of prostate cancer. *Proceedings of the National Academy of Sciences*, *101*, 811–816.
- Lin, P.-E. (1973). Procedures for testing the difference of means with incomplete data. *Journal of the American Statistical Association*, *68*, 699–703.
- Lin, P.-E., & Stivers, L. E. (1974). On difference of means with incomplete data. *Biometrika*, *61*, 325–334.
- Little, R. J. (1976). Inference about means from incomplete multivariate data. *Biometrika*, *63*, 593–604.
- Looney, S. W., & Jones, P. W. (2003). A method for comparing two normal means using combined samples of correlated and uncorrelated data. *Statistics in Medicine*, *22*, 1601–1610.
- Martínez-Cambor, P., Corral, N., & María de la Hera, J. (2013). Hypothesis test for paired samples in the presence of missing data. *Journal of Applied Statistics*, *40*, 76–87.
- Mehta, J., & Gurland, J. (1969). Testing equality of means in the presence of correlation. *Biometrika*, *56*, 119–126.
- Mehta, J., & Gurland, J. (1973). A test for equality of means in the presence of correlation and missing values. *Biometrika*, *60*, 211–213.
- Morrison, D. F. (1973). A test for equality of means of correlated variates with missing data on one response. *Biometrika*, *60*, 101–105.
- Munoz, M., Henderson, M., Haber, M., & Norris, M. (2007). Role of the MRP1/ABCC1 multidrug transporter protein in cancer. *IUBMB Life*, *59*, 752–757.
- Pauly, M., Brunner, E., & Konietschke, F. (2015). Asymptotic permutation tests in general factorial designs. *Journal of the Royal Statistical Society: Series B: Statistical Methodology*, *77*, 461–473.
- Pesarin, F. (2001). *Multivariate permutation tests: With applications in Biostatistics*. Wiley Chichester.
- Pesarin, F., & Salmaso, L. (2010). *Permutation tests for complex data: Theory, applications and software*. Wiley.
- Preece, D. (1982). T is for trouble (and textbooks): A critique of some examples of the paired-samples t-test. *Journal of the Royal Statistical Society: Series D (The Statistician)*, *31*, 169–195.
- Pruzek, R. M., & Helmreich, J. E. (2009). Enhancing dependent sample analyses with graphics. *Journal of Statistics Education*, *17*.
- Qi, Q., Yan, L., & Tian, L. (2019). Testing equality of means in partially paired data with incompleteness in single response. *Statistical Methods in Medical Research*, *28*, 1508–1522.
- Ramosaj, B., Amro, L., & Pauly, M. (2020). A cautionary tale on using imputation methods for inference in matched-pairs design. *Bioinformatics*, *36*, 3099–3106.

- Ramosaj, B., & Pauly, M. (2019). Predicting missing values: A comparative study on non-parametric approaches for imputation. *Computational Statistics*, *34*, 1741–1764.
- Rempala, G. A., & Looney, S. W. (2006). Asymptotic properties of a two sample randomized test for partially dependent data. *Journal of Statistical Planning and Inference*, *136*, 68–89.
- Rubin, D. B. (2004). *Multiple imputation for nonresponse in surveys* (Vol. 81). Wiley.
- Samawi, H. M., Helu, A., & Vogel, R. (2011). A nonparametric test of symmetry based on the overlapping coefficient. *Journal of Applied Statistics*, *38*, 885–898.
- Samawi, H. M., & Vogel, R. (2014). Notes on two sample tests for partially correlated (paired) data. *Journal of Applied Statistics*, *41*, 109–117.
- Schafer, J. L. (1999). Multiple imputation: A primer. *Statistical Methods in Medical Research*, *8*, 3–15.
- Smaga, Ł. (2017). Bootstrap methods for multivariate hypothesis testing. *Communications in Statistics-Simulation and Computation*, *46*, 7654–7667.
- Stekhoven, D. J., & Bühlmann, P. (2011). Missforest—Non-parametric missing value imputation for mixed-type data. *Bioinformatics*, *28*, 112–118.
- Sterne, J. A., White, I. R., Carlin, J. B., Spratt, M., Royston, P., Kenward, M. G., Wood, A. M., & Carpenter, J. R. (2009). Multiple imputation for missing data in epidemiological and clinical research: Potential and pitfalls. *BMJ*, *338*, b2393.
- Uddin, N., & Hasan, M. (2017). Testing equality of two normal means using combined samples of paired and unpaired data. *Communications in Statistics - Simulation and Computation*, *46*, 2430–2446.
- Vallejo, G., Fernández, M., & Livacic-Rojas, P. E. (2010). Analysis of unbalanced factorial designs with heteroscedastic data. *Journal of Statistical Computation and Simulation*, *80*, 75–88.
- Van Buuren, S. (2018). *Flexible imputation of missing data*. Chapman and Hall/CRC.
- Waljee, A. K., Mukherjee, A., Singal, A. G., Zhang, Y., Warren, J., Balis, U., Marrero, J., Zhu, J., & Higgins, P. D. (2013). Comparison of imputation methods for missing laboratory data in medicine. *BMJ Open*, *3*, e002847.
- Xu, J., & Harrar, S. W. (2012). Accurate mean comparisons for paired samples with missing data: An application to a smoking-cessation trial. *Biometrical Journal*, *54*, 281–295.
- 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.
- Yu, D., Lim, J., Liang, F., Kim, K., Kim, B. S., & Jang, W. (2012). Permutation test for incomplete paired data with application to CDNA microarray data. *Computational Statistics & Data Analysis*, *56*, 510–521.

## SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

**How to cite this article:** Amro L, Pauly M, Ramosaj B. Asymptotic-based bootstrap approach for matched pairs with missingness in a single arm. *Biometrical Journal*. 2021;63:1389–1405. <https://doi.org/10.1002/bimj.202000051>

## Supplementary Materials for: Asymptotic based bootstrap approach for matched pairs with missingness in a single-arm

Lubna Amro<sup>\*,1</sup>, Markus Pauly<sup>1</sup>, and Burim Ramosaj<sup>1</sup>

<sup>1</sup> Mathematical Statistics and Applications in Industry, Faculty of Statistics, Technical University of Dortmund, Germany

Received zzz, revised zzz, accepted zzz

In this supplementary material, we present the asymptotic distribution of the suggested bootstrapped tests, their related propositions and theorems along with their proofs. Further, we recall the definition of the different missing mechanisms and present additional type-I error and power simulation results of our suggested methods and the alternative approaches from Section 5 of the paper.

### 1 Model Assumptions and Asymptotic Analyses

First, we need to set up the following assumption regarding sample sizes, which we assume throughout

**Assumption 1** For  $\min\{n_c, n_u\} \rightarrow \infty$  we require that

- $\frac{n_c}{n_c+n_u} \rightarrow \kappa_1 \in (0, 1)$ ,
- $\frac{n_u}{n_c+n_u} \rightarrow \kappa_2 = (1 - \kappa_1) \in (0, 1)$ .

We now derive the asymptotic behavior of the statistic  $Z_n$  under the null hypothesis.

**Proposition 1.1** Let  $\sigma_1^2 = \text{Var}(X_{11}^{(c)}) = \text{Var}(X_{11}^{(i)})$ ,  $\sigma_2^2 = \text{Var}(X_{21}^{(c)})$  and  $\rho = \text{corr}(X_{11}^{(c)}, X_{21}^{(c)})$ . The statistic  $Z_n$  as given in the main article has, asymptotically, as  $n \rightarrow \infty$ , a multivariate normal distribution with expectation  $\mathbf{0}$  and covariance matrix given by

$$\Sigma = \lim_{n \rightarrow \infty} \Sigma_n = \begin{bmatrix} \kappa_1^{-1} \sigma_1^2 & \kappa_1^{-1} \rho \sigma_1 \sigma_2 & 0 \\ \kappa_1^{-1} \rho \sigma_1 \sigma_2 & \kappa_1^{-1} \sigma_2^2 & 0 \\ 0 & 0 & \kappa_2^{-1} \sigma_1^2 \end{bmatrix}. \quad (1)$$

$\Sigma_n$  can be consistently estimated by

$$\widehat{\Sigma}_n = \begin{bmatrix} \widehat{\kappa}_1^{-1} (\widehat{\sigma}_1^{(c)})^2 & \widehat{\kappa}_1^{-1} \widehat{\rho} (\widehat{\sigma}_1^{(c)})^2 \widehat{\sigma}_2 & 0 \\ \widehat{\kappa}_1^{-1} \widehat{\rho} (\widehat{\sigma}_1^{(c)})^2 \widehat{\sigma}_2 & \widehat{\kappa}_1^{-1} \widehat{\sigma}_2^2 & 0 \\ 0 & 0 & \widehat{\kappa}_2^{-1} (\widehat{\sigma}_1^{(i)})^2 \end{bmatrix}, \quad (2)$$

where  $\widehat{\kappa}_1 = n_c/n$ ,  $\widehat{\kappa}_2 = n_u/n$  and  $\widehat{\sigma}_1^{(c)}$  is the empirical standard deviation of  $\{X_{11}^{(c)}, \dots, X_{1n_c}^{(c)}\}$ ,  $\widehat{\sigma}_1^{(i)}$  is the empirical standard deviation of  $\{X_{11}^{(i)}, \dots, X_{1n_u}^{(i)}\}$ ,  $\widehat{\sigma}_2^2 = \frac{1}{n_c-1} \sum_{i=1}^{n_c} (X_{2i}^{(c)} - \bar{X}_2^{(c)})^2$  and the correlation factor  $\rho$  is estimated through the empirical correlation  $\widehat{\rho}$  calculated from the paired data  $\mathbf{X}^{(c)}$ .

\*Corresponding author: e-mail: lubna.amro@tu-dortmund.de

## 2 Asymptotic Distribution of The Quadratic Forms Statistics

**Theorem 2.1** *The statistic  $T_W$  has under the null hypothesis  $H_0 : \{\mu_1 = \mu_2\}$  and  $\Sigma > 0$ , asymptotically, as  $n \rightarrow \infty$ , a central  $\chi_2^2$  distribution.*

**Theorem 2.2** *Under the null hypothesis  $H_0 : \{\mu_1 = \mu_2\}$ , the test statistic  $T_A$  has asymptotically, as  $n \rightarrow \infty$ , the same distribution as the random variable*

$$Y = \sum_{i=1}^2 \lambda_i Y_i / \text{tr}(\mathbf{A}\Sigma\mathbf{A}^\top),$$

where  $Y_i \stackrel{i.i.d}{\sim} \chi_1^2$  and the weights  $\lambda_i$  are the eigenvalues of  $\mathbf{A}\Sigma\mathbf{A}^\top$  where  $\Sigma$  is given in (1).

**Theorem 2.3** *Under the null hypothesis  $H_0 : \{\mu_1 = \mu_2\}$  and  $\sigma_i^2 > 0$  holds for  $i=1,2$ , the test statistic  $T_M$  has asymptotically, as  $n \rightarrow \infty$ , the same distribution as the random variable*

$$\tilde{Y} = \sum_{i=1}^2 \tilde{\lambda}_i \tilde{Y}_i,$$

where  $\tilde{Y}_i \stackrel{i.i.d}{\sim} \chi_1^2$  and the weights  $\tilde{\lambda}_i$  are the eigenvalues of  $\mathbf{D}\mathbf{A}\Sigma\mathbf{A}^\top$  and  $\mathbf{D} = \text{diag}((\mathbf{A}\Sigma\mathbf{A}^\top)_{ii}^+)$ .

## 3 Asymptotic Distribution of The Parametric Bootstrapped Test

**Theorem 3.1** *For any choice  $[-] \in \{A, M, W\}$ , the conditional distribution of  $T_{[-]}^*$  converges weakly to the null distribution of  $T_{[-]}$  in probability for any choice of  $\boldsymbol{\mu} \in \mathbb{R}^2$  and  $\boldsymbol{\mu}_0 \in H_0$ . In particular we have*

$$\sup_{x \in \mathbb{R}} |P_{\boldsymbol{\mu}}(T_{[-]}^* \leq x | \mathbf{X}) - P_{\boldsymbol{\mu}_0}(T_{[-]} \leq x)| \xrightarrow{p} 0.$$

From Theorem 3.1, we thus obtain the asymptotically correct bootstrap tests  $\varphi_W^* = \mathbb{1}\{T_W > c_W^*\}$ ,  $\varphi_A^* = \mathbb{1}\{T_A > c_A^*\}$ , and  $\varphi_M^* = \mathbb{1}\{T_M > c_M^*\}$  where  $c_W^*$ ,  $c_A^*$ , and  $c_M^*$  denote the conditional  $(1 - \alpha)$ - bootstrap quantiles of  $T_W^*$ ,  $T_A^*$ , and  $T_M^*$  respectively.

## 4 Proofs

*Proof of Proposition 1.1:*

The results follow from the multivariate central limit theorem (CLT) and the law of large numbers, respectively.

*Proof of Proposition 2.1 in the paper:*

The stated convergence follows from Proposition 2.1, and an application of the continuous mapping theorem (CMT).

*Proof of Theorem 2.1:*

It follows from Proposition 2.2 that we have convergence in distribution  $\mathbf{AZ}_n \xrightarrow{d} N_2(\mathbf{0}, \mathbf{A}\Sigma\mathbf{A}^\top)$  as  $n \rightarrow \infty$  under  $H_0$ . Hence, using the CMT, the quadratic form  $\tilde{T}_W = (\mathbf{AZ}_n)^\top (\mathbf{A}\Sigma\mathbf{A}^\top)^+ (\mathbf{AZ}_n)$  has asymptotically a central  $\chi_f^2$  distribution with  $f = \text{rank}(\mathbf{A})$  degrees of freedom. Moreover, as  $\hat{\Sigma}_n$  is a consistent estimator for  $\Sigma > 0$ , the result follows from Slutsky theorem, see, e.g., Konietzschke *et al.* (2015) for similar arguments.

*Proof of Theorem 2.2:*

Applying again the CMT, it follows that  $tr(\mathbf{A}\Sigma\mathbf{A}^\top) \cdot T_A = (\mathbf{A}\mathbf{Z}_n)^T(\mathbf{A}\mathbf{Z}_n)$  has asymptotically the same distribution as  $\sum_{i=1}^2 \lambda_i Y_i$  (Graybill, 1976; Brunner and Puri, 2001). Then, the result follows from the invariance of the multivariate standard normal distribution under orthogonal transformations and the consistency of  $\widehat{\Sigma}_n$  by using Slutsky theorem.

*Proof of Theorem 2.3:*

Following similar arguments as prescribed in Friedrich and Pauly (2018), we can obtain that  $\widehat{\mathbf{D}}_n = \text{diag}((\mathbf{A}\widehat{\Sigma}_n\mathbf{A}^\top)_{ii}^+)$   $\xrightarrow{p}$   $\text{diag}((\mathbf{A}\Sigma\mathbf{A}^\top)_{ii}^+) = \mathbf{D}$  as  $\widehat{\Sigma}_n \xrightarrow{p} \Sigma > 0$ . Thus, the result follows from the representation theorem of quadratic forms (Rao et al., 1972).

*Proof of Theorem 3.1:*

First, we apply the Multivariate Lindeberg-Feller Theorem (MLFT) to show that (given the data)  $\sqrt{n}\bar{\mathbf{X}}^{*(c)}$  =  $\sqrt{n}[\bar{\mathbf{X}}_1^{*(c)}, \bar{\mathbf{X}}_2^{*(c)}]$  converges in distribution to a normal distributed random variable. We start by checking the MLFT conditions:

$$\begin{aligned} A) \sum_{k=1}^{n_c} \mathbb{E} \left( \frac{\sqrt{n}}{n_c} \mathbf{X}_k^{*(c)} | \mathbf{X} \right) &= \sum_{k=1}^{n_c} \frac{\sqrt{n}}{n_c} \mathbb{E} \left( \mathbf{X}_k^{*(c)} | \mathbf{X} \right) = 0 \\ B) \sum_{k=1}^{n_c} \text{Cov} \left( \frac{\sqrt{n}}{n_c} \mathbf{X}_k^{*(c)} | \mathbf{X} \right) &= \sum_{k=1}^{n_c} \frac{n}{n_c^2} \text{Cov} \left( \mathbf{X}_k^{*(c)} | \mathbf{X} \right) = \sum_{k=1}^{n_c} \frac{n}{n_c^2} \widehat{\Gamma} \xrightarrow{p} \frac{1}{\kappa_1} \Gamma \\ C) \lim_{n \rightarrow \infty} \sum_{k=1}^{n_c} \mathbb{E} \left( \left\| \frac{\sqrt{n}}{n_c} \mathbf{X}_k^{*(c)} \right\|^2 \cdot \mathbb{1} \left\{ \left\| \frac{\sqrt{n}}{n_c} \mathbf{X}_k^{*(c)} \right\| > \epsilon \right\} | \mathbf{X} \right) \\ &= \lim_{n \rightarrow \infty} \frac{n}{n_c^2} \sum_{k=1}^{n_c} \mathbb{E} \left( \left\| \mathbf{X}_k^{*(c)} \right\|^2 \cdot \mathbb{1} \left\{ \left\| \frac{\sqrt{n}}{n_c} \mathbf{X}_k^{*(c)} \right\| > \epsilon \right\} | \mathbf{X} \right) \\ &= \frac{1}{\kappa_1} \cdot \lim_{n \rightarrow \infty} \mathbb{E} \left( \left\| \mathbf{X}_k^{*(c)} \right\|^2 \cdot \mathbb{1} \left\{ \left\| \frac{\sqrt{n}}{n_c} \mathbf{X}_k^{*(c)} \right\| > \epsilon \right\} | \mathbf{X} \right) \\ &\leq \frac{1}{\kappa_1} \cdot \lim_{n \rightarrow \infty} \sqrt{\mathbb{E} \left( \left\| \mathbf{X}_k^{*(c)} \right\|^2 | \mathbf{X} \right)^2} \cdot \sqrt{\mathbb{E} \left( \mathbb{1} \left\{ \left\| \frac{\sqrt{n}}{n_c} \mathbf{X}_k^{*(c)} \right\| > \epsilon \right\} | \mathbf{X} \right)^2} \end{aligned}$$

The last step follows from the Cauchy-Schwarz inequality. Now, the first term  $\mathbb{E} \left( \left\| \mathbf{X}_k^{*(c)} \right\|^2 | \mathbf{X} \right)^2$  is asymptotically bounded, while the second term converges to zero in probability since  $\mathbb{1} \left\{ \left\| \frac{\sqrt{n}}{n_c} \mathbf{X}_k^{*(c)} \right\| > \epsilon \right\} = 1$  holds iff  $\left\| \mathbf{X}_k^{*(c)} \right\| > \frac{n_c}{\sqrt{n}} \epsilon = \frac{n_c}{n} \sqrt{n} \epsilon$ . As  $\frac{n_c}{n} \sqrt{n} \epsilon \rightarrow \infty$  while  $\mathbf{X}_k^{*(c)} \xrightarrow{d} N(0, \Gamma)$ , it follows that the Lindeberg condition is satisfied (in probability). Thus, proves that the conditional distribution of  $\sqrt{n}\bar{\mathbf{X}}^{*(c)}$  given the data weakly converges to  $\frac{1}{\kappa_1} N(0, \Gamma)$  in probability.

In a similar way, we proof that  $\sqrt{n}\bar{\mathbf{X}}_1^{*(i)}$  given the data weakly converges to  $\frac{1}{\kappa_2} N(0, \sigma_1^2)$  in probability.

Now, due to the MCAR setting,  $\mathbf{X}^c$  is independent of  $\mathbf{X}^{(i)}$  and by using Slutsky,  $\mathbf{Z}_n^* = \sqrt{n}[\bar{\mathbf{X}}_1^{*(c)}, \bar{\mathbf{X}}_2^{*(c)}, \bar{\mathbf{X}}_1^{*(i)}]^\top$  converges in distribution to  $N_3(0, \Sigma)$  where,  $\Sigma$  is defined in Section 2 in the paper. Following the same steps as in the proof of Theorem 3.1-3.3, this concludes the proof.

### 5 MCAR, MAR and MNAR

To explain the different missing schemes we define a missing indicator variable  $\mathbf{R}_j = [R_{1j}, R_{2j}]^\top \in \mathbb{R}^2$ , that identify what is known and what is missing, i.e.  $R_{ij} = 0$  if  $X_{ij}$  is missing and  $R_{ij} = 1$  otherwise,

$j = 1, \dots, n$ . Then, Rubin defines the missing mechanism through a parametric distributional model on  $\mathbf{R} = \{\mathbf{R}_j\}_{j=1}^n$  and classifies their presence through Missing Completely at Random (MCAR), Missing at Random (MAR) and Missing not at Random (MNAR) schemes (Rubin, 2004). To describe this in our model let us denote the observed data as  $\mathbf{X}^{obs} = (\mathbf{X}^{(c)}, \mathbf{X}^{(i)})$  and denote with  $\mathbf{X}^{mis}$  the missing observations.

The data are said to be **MCAR** if the probability of an observation being missing does not depend on observed or unobserved data, i.e. if  $P(\mathbf{R} | \mathbf{X}^{obs}, \mathbf{X}^{mis}) = P(\mathbf{R})$ .

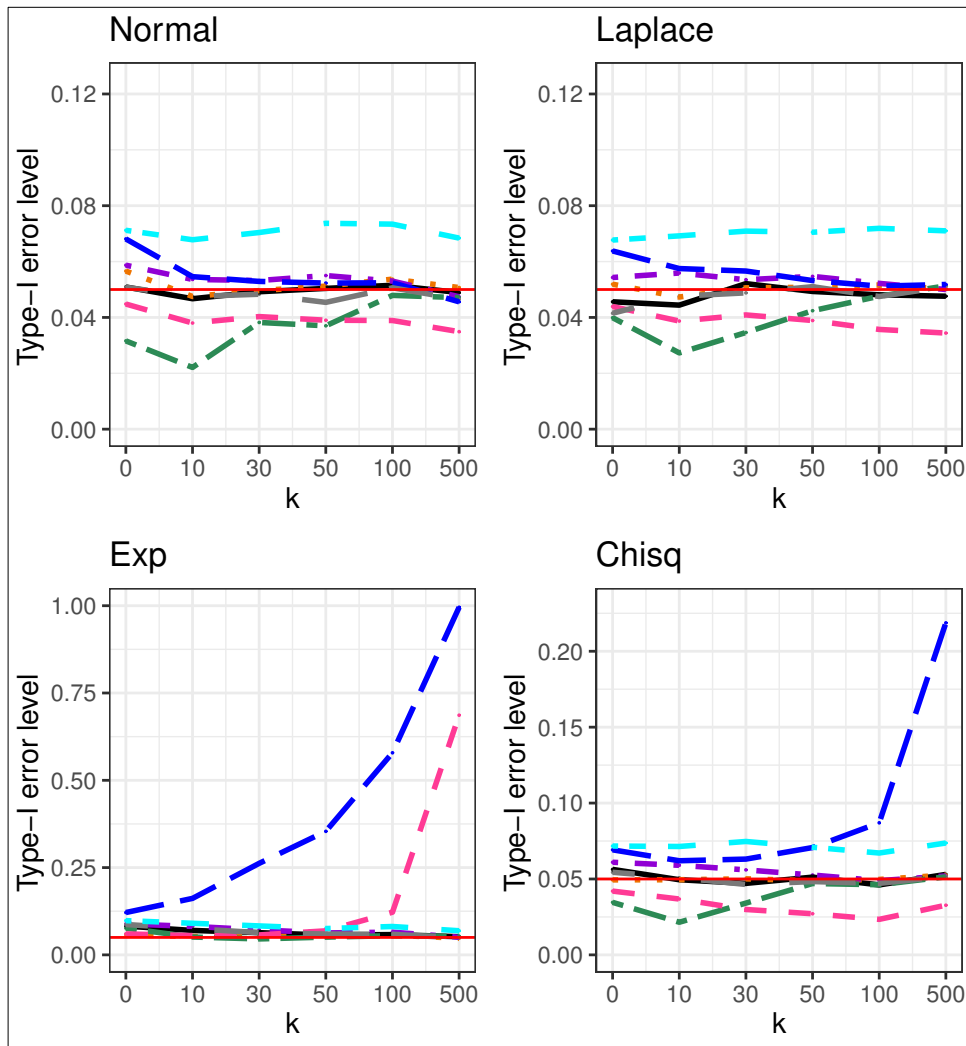
Data are said to be **MAR** if the probability of missingness may depend on observed data but does not depend on unobserved data,  $P(\mathbf{R} | \mathbf{X}^{obs}, \mathbf{X}^{mis}) = P(\mathbf{R} | \mathbf{X}^{obs})$ .

Finally, data are said to be **MNAR**, if missingness does depend on the unobserved data. It can easily be seen that MAR includes MCAR as a special case. For more details about the different missing mechanisms, we refer to the monograph of Little and Rubin (2014).

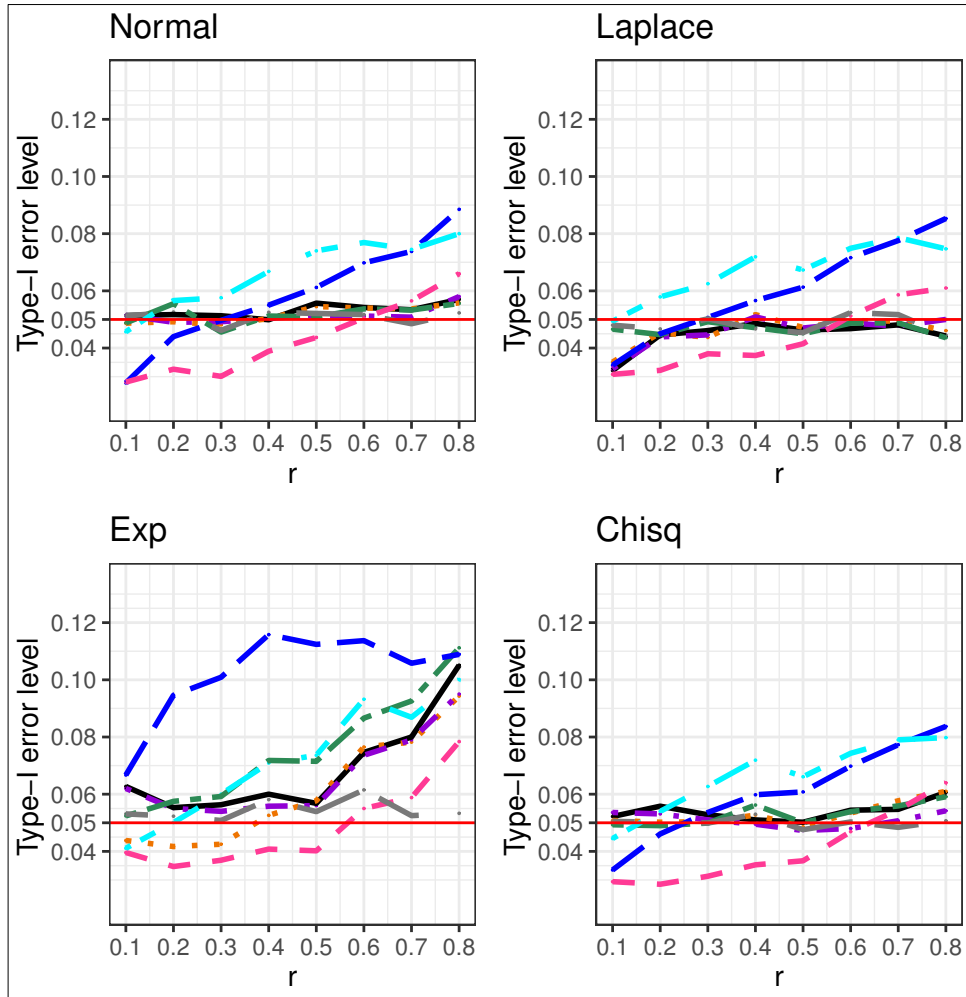


### 6 Type-I Error and Power Results

In the sequel, we present some additional type-I error and power results of the Monte Carlo simulation study, that is described in detail in Section 6 of the paper, for testing  $H_0$  for matched pairs with missingness in one arm under the MCAR, and MAR schemes.



**Figure 1** Type-I error simulation results ( $\alpha = 0.05$ ) of the tests  $T_W^*$  (—),  $T_A^*$  (· · ·),  $T_M^*$  (- · -),  $T_L^*$  (- - -),  $T_N$  (— —),  $T_F$  (- - -),  $T_t$  (— —), and  $T_P$  (- · ·) for different distributions under correlation value ( $\rho = 0.9$ ) and heteroscedastic covariance matrix  $\Sigma_2$  for varying  $k$  values added to  $(n_c, n_u) = (10, 10)$  under the MCAR framework.



**Figure 2** Type-I error simulation results ( $\alpha = 0.05$ ) of the tests  $T_W^*$  (—),  $T_A^*$  (····),  $T_M^*$  (- · - ·),  $T_L^*$  (- - -),  $T_N^*$  (— —),  $T_F^*$  (- - -),  $T_t$  (— —), and  $T_P$  (- · - ·) for different distributions under correlation value ( $\rho = 0.5$ ) with sample size ( $n = 30$ ) and heteroscedastic covariance matrix  $\Sigma_2$  for varying missing rates  $r \in \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8\}$  under the MCAR framework.

**Table 1** Type-I error simulation results ( $\alpha = 0.05$ ) of the tests for different distributions under varying correlation values ( $\rho$ ) with sample sizes  $(n_c, n_u) = (10, 30)$  and different covariance matrices  $\Sigma_1$  and  $\Sigma_2$  under the MCAR framework. For each setting, the values closest to the prescribed level are printed in **bold** and values exceeding the upper limit (6.8%) of the 99% binomial interval are in **red** colour.

Dist	$\rho$	$\Sigma_1$								$\Sigma_2$									
		Parametric bootstrap				Alternatives				Parametric bootstrap				Alternatives					
		$F$	$T_W^*$	$T_A^*$	$T_M^*$	$T_L^*$	$T_t$	$T_N$	$T_F$	$T_P$	$F$	$T_W^*$	$T_A^*$	$T_M^*$	$T_L^*$	$T_t$	$T_N$	$T_F$	$T_P$
Normal	-0.9	<b>5.1</b>	<b>5.1</b>	5.5	<b>6.9</b>	<b>5.1</b>	5.2	<b>7.2</b>	5.4	<b>8.6</b>	5.2	<b>5.1</b>	5.4	<b>7.2</b>	<b>5.1</b>	5.3	<b>8.1</b>	6.7	<b>9.4</b>
	-0.5	5.4	5.4	5.4	6	5.5	<b>4.7</b>	6.2	4.5	<b>7.5</b>	5.4	5.4	<b>5.3</b>	6.3	5.6	<b>4.7</b>	<b>6.9</b>	5.8	<b>8.3</b>
	-0.1	5.4	<b>5.1</b>	5.8	6.1	5.6	<b>5.1</b>	<b>7</b>	4.8	<b>7.5</b>	5.3	<b>5.1</b>	5.8	6.6	5.6	5.2	<b>7.8</b>	6.2	<b>8.6</b>
	0.1	5.5	5.4	5.7	6.1	5.6	<b>5.3</b>	<b>6.8</b>	<b>4.7</b>	<b>7.4</b>	5.4	5.4	6	6.5	5.6	<b>5.2</b>	<b>7.5</b>	6.2	<b>8.4</b>
	0.5	5.6	5.4	5.4	5.5	5.9	<b>5.3</b>	6.2	4.4	6.2	5.5	5.5	5.9	6.2	6.1	<b>5.4</b>	<b>7.5</b>	6.3	<b>8.1</b>
	0.9	<b>5.2</b>	5.8	5.6	4.5	3.6	5.3	5.8	4.4	5.6	5.3	<b>5.2</b>	5.8	<b>7.1</b>	3.1	<b>4.8</b>	<b>7.4</b>	6	<b>7.8</b>
Laplace	-0.9	<b>5.0</b>	4.8	5.4	<b>7</b>	4.8	5.3	<b>7.2</b>	5.4	<b>8.6</b>	<b>5.0</b>	4.8	5.3	<b>7.1</b>	4.9	5.1	<b>8</b>	6.4	<b>9.5</b>
	-0.5	<b>4.8</b>	4	4.6	<b>5.2</b>	4.4	4.3	6.2	4.4	<b>7.3</b>	<b>4.9</b>	3.9	4.5	5.7	4.4	4.2	<b>7</b>	5.5	<b>8.1</b>
	-0.1	5.4	4.5	<b>5.1</b>	5.5	4.8	4.4	6.6	4.2	<b>6.9</b>	5.4	4.4	<b>5.0</b>	5.8	4.9	4.4	<b>7.3</b>	5.4	<b>7.8</b>
	0.1	4.7	4.3	<b>5.0</b>	5.4	4.8	4.6	6.5	4	<b>6.8</b>	4.7	4.5	4.7	5.8	<b>4.9</b>	4.5	<b>7</b>	5.2	<b>7.7</b>
	0.5	5.5	4.4	4.7	<b>5.1</b>	4.8	4.5	6.4	4.2	5.9	<b>5.2</b>	4.4	<b>4.8</b>	5.7	4.6	4.4	<b>6.8</b>	<b>5.2</b>	<b>7.2</b>
	0.9	5.3	4.6	5.2	4.2	3.4	4.9	6.1	4.4	<b>5.0</b>	<b>5.0</b>	4.6	5.2	6.6	3	4.3	<b>6.8</b>	5.5	<b>7</b>
Exponential	-0.9	4.9	4.9	4.7	6.7	5.4	4.3	<b>7.1</b>	<b>5.0</b>	<b>8.3</b>	5.2	5.4	5.9	<b>8.1</b>	6.5	<b>5.1</b>	10.1	5.8	<b>10</b>
	-0.5	<b>4.9</b>	5.7	<b>5.1</b>	<b>6.8</b>	6.7	4.5	6.7	4.6	<b>7.4</b>	5.3	6.4	6.4	<b>8.1</b>	<b>8</b>	<b>4.8</b>	10.1	5.9	<b>9.7</b>
	-0.1	<b>5.1</b>	<b>7.3</b>	5.9	<b>7.5</b>	<b>7.8</b>	4.5	6.7	4.8	<b>7.3</b>	5.4	<b>8.1</b>	<b>7.7</b>	<b>9</b>	<b>9</b>	<b>5.1</b>	11.3	6.6	<b>9.8</b>
	0.1	<b>5.2</b>	<b>7.8</b>	6.3	<b>7.4</b>	<b>8.6</b>	4.1	6.7	4.5	<b>7.2</b>	5.7	<b>8.7</b>	<b>8.2</b>	<b>9.2</b>	<b>9.7</b>	<b>5.3</b>	11.3	<b>7</b>	<b>10</b>
	0.5	<b>4.8</b>	<b>7.9</b>	5.6	6.3	<b>9</b>	4.1	6.2	4.4	5.9	<b>5.2</b>	<b>8.5</b>	<b>8.1</b>	<b>8.6</b>	<b>9.8</b>	5.3	10.8	6	<b>8.9</b>
	0.9	<b>5.1</b>	<b>9.5</b>	6.7	4.8	<b>9.7</b>	4.2	6.3	4.8	<b>4.9</b>	<b>6.9</b>	<b>9.6</b>	<b>8.4</b>	<b>11.5</b>	<b>7.9</b>	<b>8.9</b>	<b>13.1</b>	<b>6.8</b>	<b>11.2</b>
Chisquare	-0.9	<b>5.1</b>	<b>5.1</b>	5.5	<b>6.8</b>	4.6	<b>5.1</b>	<b>6.9</b>	<b>4.9</b>	<b>8.6</b>	<b>5.0</b>	5.1	5.4	<b>7.3</b>	4.9	5.1	<b>8.1</b>	6	<b>9.6</b>
	-0.5	<b>5.0</b>	5.1	5.7	6.7	5.2	<b>5.0</b>	<b>7.2</b>	<b>5.0</b>	<b>8.2</b>	<b>5.1</b>	5.2	5.7	<b>7.2</b>	5.3	<b>4.9</b>	<b>7.8</b>	6.1	<b>9</b>
	-0.1	<b>5.1</b>	5.2	5.5	6	5.7	4.8	6.2	4.4	<b>7.1</b>	<b>4.9</b>	5.3	5.5	6.2	5.8	4.8	<b>7.5</b>	5.6	<b>8.1</b>
	0.1	<b>4.9</b>	5.5	5.2	5.8	5.3	<b>4.9</b>	6.2	4.3	6.6	<b>4.9</b>	5.6	5.4	6.3	5.6	<b>4.9</b>	<b>7.2</b>	5.5	<b>8.1</b>
	0.5	5.4	5.5	<b>5.1</b>	5.2	5.3	4.8	6	4.2	5.9	5.5	5.6	5.7	6	5.6	<b>5.1</b>	<b>7.2</b>	5.6	<b>7.7</b>
	0.9	4.8	5.7	5.6	4.1	3.9	4.6	5.9	4.4	<b>5.1</b>	<b>5.0</b>	5.6	6.1	<b>7.4</b>	3.3	5.1	<b>7.6</b>	5.6	<b>8</b>

**Table 2** Type-I error simulation results ( $\alpha = 0.05$ ) of the tests for different distributions under varying correlation values ( $\rho$ ) with sample sizes  $(n_c, n_u) = (30, 10)$  and different covariance matrices  $\Sigma_1$  and  $\Sigma_2$  under the MCAR framework. For each setting, the values closest to the prescribed level are printed in **bold** and values exceeding the upper limit (6.8%) of the 99% binomial interval are in **red** colour.

Dist	$\rho$	$\Sigma_1$								$\Sigma_2$									
		$F$	Parametric bootstrap				Alternatives				$F$	Parametric bootstrap				Alternatives			
			$T_W^*$	$T_A^*$	$T_M^*$	$T_L^*$	$T_t$	$T_N$	$T_F$	$T_P$		$T_W^*$	$T_A^*$	$T_M^*$	$T_L^*$	$T_t$	$T_N$	$T_F$	$T_P$
Normal	-0.9	4.9	<b>5.0</b>	4.8	4.9	4.9	4.6	5.2	3.5	6	<b>4.9</b>	5.2	<b>4.9</b>	4.7	<b>4.9</b>	4.7	4.6	3	6.6
	-0.5	4.8	4.8	4.4	4.7	4.8	<b>4.9</b>	<b>4.9</b>	3.5	5.9	<b>4.8</b>	<b>4.8</b>	4.6	4.6	<b>4.8</b>	<b>4.8</b>	4.1	3	6.2
	-0.1	4.9	<b>5.0</b>	4.7	4.7	4.9	4.7	5.3	3.6	5.5	<b>4.9</b>	<b>4.9</b>	<b>4.9</b>	4.8	4.8	4.7	4.4	3.1	5.9
	0.1	5.3	4.9	<b>5.0</b>	4.9	5.3	5.2	5.5	3.5	5.7	5.1	4.8	5.1	<b>5.0</b>	5.2	<b>5.0</b>	4.4	3	6.3
	0.5	5.5	<b>5.0</b>	4.6	4.7	4.7	4.7	5.2	3.7	5.2	5.4	<b>4.9</b>	4.7	4.6	4.7	4.6	4.4	2.8	5.7
	0.9	5.6	5.3	5.2	5.1	2.3	<b>5.0</b>	5.3	4.4	5.2	5.7	5.4	<b>5.1</b>	5.3	2	5.3	4.8	3.3	6.3
Laplace	-0.9	<b>5.0</b>	4.7	<b>5.0</b>	5.3	4.9	4.8	5.4	3.6	6.4	<b>5.0</b>	4.7	<b>5.0</b>	5.4	4.9	4.9	4.8	3.5	<b>6.9</b>
	-0.5	5.2	4.6	5.1	<b>5.0</b>	5.3	4.7	5.4	3.8	6.3	5.2	4.7	<b>5.1</b>	5.2	5.2	4.8	4.8	3.5	6.7
	-0.1	5.2	4.5	<b>5.0</b>	4.8	4.9	4.8	5.4	3.9	6	5.2	4.6	4.9	5.1	5.1	4.9	<b>5.0</b>	3.5	6.5
	0.1	5.5	<b>4.9</b>	4.8	5.2	5.6	5.4	5.5	4.3	6.4	5.7	<b>5.0</b>	<b>5.0</b>	5.4	5.8	5.5	5.3	3.7	6.7
	0.5	<b>5.0</b>	4.5	4.2	4.5	4.8	4.8	5.1	3.6	5.3	<b>4.8</b>	4.7	4.4	4.6	4.8	4.7	4.6	3.1	5.6
	0.9	4.9	4.5	4.6	4.3	3.3	<b>5.0</b>	5.1	4.4	<b>5.0</b>	5.2	4.7	4.7	<b>5.1</b>	2.6	5.3	4.6	3.6	6.3
Exponential	-0.9	5.1	5.1	<b>5.0</b>	5.9	5.1	4.7	5.3	3.7	6.6	5.3	<b>4.9</b>	5.2	6	5.2	<b>5.1</b>	<b>9.1</b>	2.8	<b>6.8</b>
	-0.5	<b>4.7</b>	6.4	4.4	6.3	<b>4.7</b>	4.4	<b>5.3</b>	3.5	5.8	<b>4.9</b>	5.8	4.5	5.7	<b>5.1</b>	4.6	<b>9.2</b>	2.5	5.9
	-0.1	<b>5.0</b>	<b>6.8</b>	4.7	6.2	4.6	4.5	5.1	3.7	5.6	5.2	6.4	4.6	5.7	<b>5.0</b>	4.8	<b>10.7</b>	2.7	5.6
	0.1	<b>5.1</b>	6.5	<b>4.9</b>	6.2	5.4	<b>5.1</b>	5.4	3.9	5.9	<b>5.3</b>	6.2	<b>4.7</b>	5.9	5.8	5.4	<b>11</b>	3.4	6.1
	0.5	<b>4.9</b>	6	<b>5.1</b>	5.8	5.2	<b>4.9</b>	5.2	4.1	5.8	<b>5.4</b>	6.1	4.5	6	6.2	5.5	<b>12.1</b>	3.9	5.9
	0.9	5.5	5.5	5.2	5.1	4.2	<b>5.0</b>	5.2	4.7	4.9	<b>6.8</b>	<b>7</b>	<b>4.6</b>	<b>7.1</b>	4.3	6.5	<b>17</b>	6	<b>6.8</b>
Chisquare	-0.9	5.4	5.1	<b>5.0</b>	<b>5.0</b>	5.3	4.6	5.8	3.7	6.1	5.4	<b>5.0</b>	5.2	<b>5.0</b>	5.3	4.7	5.1	3	<b>6.9</b>
	-0.5	5.3	5.3	<b>5.1</b>	<b>5.1</b>	5.3	<b>5.1</b>	5.4	3.6	6.1	5.3	5.1	5.1	<b>5.0</b>	5.2	5.2	<b>5.0</b>	3	6.6
	-0.1	5.2	5.3	<b>4.9</b>	<b>5.1</b>	5.2	<b>4.9</b>	4.8	3.5	5.8	5.1	5.2	4.9	5.1	5.2	<b>5.0</b>	4.7	2.5	6.2
	0.1	5.1	5.4	<b>5.0</b>	5.1	<b>5.0</b>	4.9	5.1	3.9	5.5	5.3	5.2	<b>5.0</b>	<b>5.0</b>	<b>5.0</b>	4.9	4.8	2.8	5.9
	0.5	<b>5.0</b>	5.3	5.2	4.9	4.7	4.8	<b>5.0</b>	3.7	5.5	5.1	5.2	<b>5.0</b>	5.2	4.9	<b>5.0</b>	4.8	2.5	5.9
	0.9	<b>5.0</b>	4.9	4.6	4.6	2.5	4.9	<b>5.0</b>	4.3	5.1	5.2	4.7	4.6	<b>5.0</b>	1.6	4.7	4.8	2.8	5.8

**Table 3** Type-I error simulation results ( $\alpha = 0.05$ ) of the tests for different distributions under varying correlation values ( $\rho$ ), with sample sizes  $n = 10$  (numbers of subjects) and different covariance matrices  $\Sigma_1$  and  $\Sigma_2$  under the MAR framework. For each setting, the values closest to the prescribed level are printed in **bold** and values exceeding the upper limit (6.8%) of the 99% binomial interval are in **red** colour.

Dist	$\rho$	$\Sigma_1$								$\Sigma_2$									
		F	Parametric bootstrap				Alternatives				F	Parametric bootstrap				Alternatives			
			$T_W^*$	$T_A^*$	$T_M^*$	$T_L^*$	$T_t$	$T_N$	$T_F$	$T_P$		$T_W^*$	$T_A^*$	$T_M^*$	$T_L^*$	$T_t$	$T_N$	$T_F$	$T_P$
Normal	-0.9	4.6	5.7	<b>4.8</b>	5.3	4.5	4.7	<b>7.8</b>	3.5	6.2	5.3	5.6	5.7	5.7	<b>5.0</b>	5.4	<b>8</b>	4.1	<b>7.1</b>
	-0.5	<b>4.9</b>	4.7	4.5	<b>4.9</b>	4.2	4.4	<b>7.2</b>	3.6	6	5.2	<b>5.1</b>	<b>5.1</b>	<b>5.1</b>	4.4	4.8	<b>7.2</b>	3.7	6.5
	-0.1	5.1	4.7	4.2	4.5	4.2	<b>5.0</b>	<b>7.2</b>	3.4	5.9	5.2	<b>5.1</b>	4.5	4.4	4.3	<b>5.1</b>	<b>7</b>	3.6	6
	0.1	<b>5.1</b>	4.8	4.2	4.3	3.8	<b>5.1</b>	<b>7</b>	3	5.8	4.9	<b>5.0</b>	4.3	4.5	3.6	4.8	<b>6.8</b>	3.1	5.9
	0.5	4.8	4.6	4.2	4.3	3.1	4.7	6.3	2.9	<b>5.1</b>	<b>5.0</b>	4.9	4.5	4.6	3.4	5.2	6.4	3.5	5.7
	0.9	4.8	4.6	<b>4.9</b>	3.8	<b>7.8</b>	5.2	5.9	2.9	5.2	4.5	4.8	4.7	<b>5.0</b>	<b>6.9</b>	5.1	<b>6.8</b>	3.3	5.8
Laplace	-0.9	4.6	4.3	4.4	<b>5.3</b>	4.3	4	<b>7.8</b>	3.6	6.5	<b>4.7</b>	4.2	4.6	6	4.4	4.2	<b>7.8</b>	3.8	6.7
	-0.5	<b>5.0</b>	3.6	4.5	4.4	4.2	4.4	<b>7.6</b>	3.8	<b>6.8</b>	<b>4.6</b>	3.7	4.3	<b>4.6</b>	3.7	4.2	<b>7</b>	3.5	6.7
	-0.1	<b>4.4</b>	3.3	3.6	3.2	3.2	4.2	<b>6.9</b>	3.3	6	<b>5.0</b>	3.2	3.8	3.8	3.2	4.2	<b>6.9</b>	3.4	6.4
	0.1	<b>5.1</b>	3.7	3.3	3.4	3.4	4.5	<b>7</b>	3.5	6	<b>5.1</b>	3.7	3.6	3.9	3.3	4.7	<b>6.9</b>	3.6	6.1
	0.5	<b>4.5</b>	3.6	3.1	3.1	3.1	4.4	6.1	3.1	5.7	<b>4.7</b>	3.5	3.5	3.3	2.7	4.4	6.7	3.1	5.7
	0.9	<b>5.0</b>	3.5	4	3.2	<b>7.7</b>	4.6	5.2	2.7	4.7	<b>4.8</b>	3.7	4	4.4	6.4	4.1	<b>6.9</b>	3.2	5.6
Exponential	-0.9	4.4	4	3.9	<b>5.5</b>	4.1	3.7	<b>7.7</b>	3.8	6.3	<b>5.1</b>	3.6	<b>5.1</b>	6.2	4.8	4.7	<b>9.2</b>	4.2	<b>7.2</b>
	-0.5	4.5	<b>5.1</b>	3.9	<b>4.9</b>	4	3.8	<b>7.7</b>	3.9	<b>6.8</b>	<b>4.6</b>	4	4	4.3	<b>4.6</b>	4.2	<b>7.8</b>	3.8	6.3
	-0.1	4.3	<b>5.1</b>	3.4	4.5	3.8	3.8	<b>7.1</b>	3.7	6.3	4.6	4.4	3.7	3.9	<b>4.7</b>	4.5	<b>7.9</b>	4	6.1
	0.1	4	<b>4.6</b>	2.7	4	3.7	4	<b>7</b>	3.4	5.8	4.9	4.5	3.4	4	<b>5.0</b>	4.7	<b>8.2</b>	4.2	6
	0.5	4.5	3.7	3.2	3.5	4.6	3.7	6.5	3.2	<b>5.2</b>	5.9	<b>5.0</b>	3.1	4.4	5.6	5.9	<b>8.2</b>	4.4	6.4
	0.9	4.5	4	4	3.8	<b>8.1</b>	3.6	5.9	2.8	<b>4.7</b>	<b>9.3</b>	<b>7.6</b>	3.6	<b>7.5</b>	<b>10.5</b>	<b>9.4</b>	<b>8.6</b>	<b>5.0</b>	<b>8.8</b>
Chisquare	-0.9	5.3	5.8	5.3	5.9	<b>5.1</b>	<b>4.9</b>	<b>7.9</b>	3.8	<b>6.8</b>	5.7	5.5	5.9	6.1	<b>5.3</b>	5.4	<b>8.3</b>	4.2	<b>7.4</b>
	-0.5	<b>5.0</b>	5.3	4.5	4.7	3.9	4.8	<b>6.9</b>	3.6	6.1	4.7	<b>5.0</b>	4.9	5.1	4.3	4.9	<b>7.5</b>	3.7	6.3
	-0.1	5.3	<b>4.9</b>	4.7	4.7	4	<b>5.1</b>	<b>7.3</b>	3.6	6.2	5.3	<b>5.0</b>	4.9	4.7	4.2	5.2	<b>7.4</b>	3.6	6.2
	0.1	5.4	5.3	4.5	4.6	3.9	<b>5.1</b>	<b>7.4</b>	3.6	6.1	5.1	<b>5.0</b>	4.9	4.9	4.2	5.2	<b>7.2</b>	3.8	6.3
	0.5	5.4	4.6	4.1	4.2	3.3	<b>4.8</b>	6.5	3.1	5.3	5.5	4.7	4.1	4.2	3.4	<b>5.2</b>	6.5	3.2	5.8
	0.9	5.4	4.6	4.3	3.8	<b>8.1</b>	4.9	5.5	2.5	<b>5.0</b>	5.5	4.5	4.7	<b>5.2</b>	<b>7.4</b>	5.5	<b>7</b>	3.5	6.2

**Table 4** Type-I error simulation results ( $\alpha = 0.05$ ) of the tests for different distributions under varying correlation values ( $\rho$ ), with sample sizes  $n = 20$  (numbers of subjects) and different covariance matrices  $\Sigma_1$  and  $\Sigma_2$  under the MAR framework. For each setting, the values closest to the prescribed level are printed in **bold** and values exceeding the upper limit (6.8%) of the 99% binomial interval are in **red** colour.

Dist	$\rho$	$\Sigma_1$								$\Sigma_2$									
		$F$	Parametric bootstrap				Alternatives				$F$	Parametric bootstrap				Alternatives			
			$T_W^*$	$T_A^*$	$T_M^*$	$T_L^*$	$T_t$	$T_N$	$T_F$	$T_P$		$T_W^*$	$T_A^*$	$T_M^*$	$T_L^*$	$T_t$	$T_N$	$T_F$	$T_P$
Normal	-0.9	5.7	5.8	5.9	6.4	<b>5.3</b>	5.5	6	4.1	<b>7.3</b>	<b>4.9</b>	5.3	5.4	5.9	4.6	<b>5.1</b>	5.3	3.8	<b>7.1</b>
	-0.5	5.4	5.4	<b>5.1</b>	5.7	4.8	<b>4.9</b>	5.8	3.8	6.4	5.2	5.1	5.1	5.6	<b>5.0</b>	5.1	5.1	3.5	6.7
	-0.1	<b>4.9</b>	5.4	4.7	5.2	4.7	<b>4.9</b>	5.8	3.8	6	5.4	4.8	<b>4.9</b>	5.2	<b>4.9</b>	<b>4.9</b>	<b>5.1</b>	3.5	6.3
	0.1	<b>5.0</b>	<b>5.0</b>	4.7	5.1	4.8	<b>5.0</b>	5.7	3.8	5.9	5.3	5.4	<b>5.1</b>	5.4	5.4	5.3	5.3	3.7	6.3
	0.5	5.5	<b>4.9</b>	4.5	4.8	4.5	5.2	5.4	4.1	5.8	<b>5.1</b>	4.8	4.6	<b>4.9</b>	4.4	<b>4.9</b>	<b>5.1</b>	3.3	5.8
	0.9	5.4	4.7	5.2	4.5	4.2	<b>5.1</b>	5.2	4	5.4	<b>5.0</b>	<b>5.0</b>	5.2	5.5	3.1	5.2	5.4	3.4	6.3
Laplace	-0.9	4.7	<b>4.8</b>	<b>4.8</b>	6.3	4.4	<b>4.8</b>	5.4	3.8	6.4	5.2	4.2	5.4	6.7	<b>4.9</b>	4.7	5.8	3.9	<b>7.5</b>
	-0.5	5.3	4.3	4.7	5.6	<b>5.0</b>	<b>5.0</b>	5.6	3.7	6.6	5.2	3.9	<b>5.0</b>	6.1	4.8	4.6	5.4	3.5	<b>7</b>
	-0.1	<b>4.8</b>	4	4.1	4.6	4.6	<b>4.8</b>	5.6	3.9	5.9	4.9	4	4.3	5.3	4.5	4.5	<b>5.0</b>	3.4	6.4
	0.1	<b>4.9</b>	3.7	3.6	4.2	4.2	4.4	5.4	3.6	5.6	5.1	4.2	4.6	5.3	5.1	<b>5.0</b>	5.4	3.7	<b>6.9</b>
	0.5	<b>5.0</b>	4.3	4.1	4.2	3.9	4.6	5.1	3.9	5.5	4.6	4.2	3.8	4.5	4.5	<b>5.2</b>	<b>5.2</b>	3.4	5.9
	0.9	5.4	4.4	4.3	4.2	4.4	4.8	<b>5.1</b>	4.2	<b>5.1</b>	<b>4.9</b>	4.1	4.7	<b>4.9</b>	3.4	4.5	5.3	3.6	5.9
Exponential	-0.9	<b>5.0</b>	5.6	4.9	<b>7.3</b>	4.8	4.8	5.6	4	<b>7.1</b>	5.3	4.7	5.6	<b>8.5</b>	<b>5.0</b>	5.2	<b>7.2</b>	3.4	<b>7.5</b>
	-0.5	4.8	<b>7.7</b>	4.5	<b>7</b>	<b>5.0</b>	4.4	5.9	4	6.6	4.7	6.7	4.7	6.4	<b>5.0</b>	4.5	<b>7</b>	3.2	6
	-0.1	<b>4.7</b>	<b>7.5</b>	4	6.3	<b>4.7</b>	4.4	6	4	6.2	<b>4.8</b>	<b>7.3</b>	4.6	6	5.6	5.3	<b>7.7</b>	3.5	6.1
	0.1	<b>4.8</b>	<b>7.1</b>	4	6.4	4.2	4.3	5.5	3.9	6.3	<b>5.0</b>	6.4	4.4	5.8	5.9	5.4	<b>7.3</b>	3.3	6.1
	0.5	<b>5.1</b>	5.6	4.8	5.4	4	4.7	5.8	4.2	6	6	6.6	<b>4.1</b>	6	<b>5.9</b>	<b>5.9</b>	<b>7.9</b>	3.9	6.4
	0.9	<b>5.2</b>	5.4	6.3	5.6	5.3	<b>4.8</b>	5.5	<b>4.8</b>	5.8	<b>7.5</b>	<b>7.7</b>	4.4	<b>8</b>	6.3	<b>7.7</b>	<b>8.8</b>	<b>4.5</b>	<b>7.5</b>
Chisquare	-0.9	<b>5.1</b>	5.5	5.2	5.8	<b>4.9</b>	5.3	5.5	3.5	<b>6.8</b>	4.9	5.2	<b>5.0</b>	5.9	4.6	4.7	5.1	3.1	6.5
	-0.5	5.3	5.8	<b>5.0</b>	5.5	5.2	<b>5.0</b>	5.8	4	6.2	4.8	5.2	<b>5.0</b>	5.6	4.9	4.7	5.1	3.2	6.7
	-0.1	4.7	5.3	4.7	<b>5.1</b>	4.8	<b>4.9</b>	5.6	3.8	5.9	<b>5.1</b>	5.5	5.2	5.7	5.3	<b>5.1</b>	5.3	3.8	6.6
	0.1	<b>5.1</b>	<b>5.1</b>	4.7	<b>4.9</b>	4.5	4.5	5.5	3.6	5.7	5.4	5.5	<b>5.0</b>	5.3	5.1	<b>5.0</b>	5.3	3.3	6.4
	0.5	5.3	5.3	<b>4.9</b>	<b>5.1</b>	4.1	<b>4.9</b>	<b>5.1</b>	3.7	5.6	5.3	<b>5.1</b>	4.7	<b>5.1</b>	4.4	<b>4.9</b>	5.2	3.1	5.9
	0.9	5.2	<b>5.1</b>	<b>5.1</b>	4.6	4.2	<b>5.1</b>	5.3	4.2	5.4	5.6	5.2	<b>4.9</b>	5.7	3.8	5.2	5.5	3.3	6.4

**Table 5** Type-I error simulation results ( $\alpha = 0.05$ ) of the tests for different distributions under varying correlation values ( $\rho$ ), with sample sizes  $n = 30$  (numbers of subjects) and different covariance matrices  $\Sigma_1$  and  $\Sigma_2$  under the MAR framework. For each setting, the values closest to the prescribed level are printed in **bold** and values exceeding the upper limit (6.8%) of the 99% binomial interval are in **red** colour.

Dist	$\rho$	$\Sigma_1$								$\Sigma_2$									
		F	Parametric bootstrap				Alternatives				F	Parametric bootstrap				Alternatives			
			$T_W^*$	$T_A^*$	$T_M^*$	$T_L^*$	$T_t$	$T_N$	$T_F$	$T_P$		$T_W^*$	$T_A^*$	$T_M^*$	$T_L^*$	$T_t$	$T_N$	$T_F$	$T_P$
Normal	-0.9	5.4	<b>5.0</b>	5.3	5.7	4.8	5.2	5.4	3.5	<b>6.8</b>	5.3	5.2	5.4	6.4	5.3	4.8	<b>5.0</b>	3.6	<b>7.5</b>
	-0.5	5.1	4.8	4.9	5.3	<b>5.0</b>	5.1	5.4	3.4	6.2	5.2	5.3	5.2	5.5	<b>5.1</b>	5.2	4.6	3.4	6.7
	-0.1	5.2	<b>5.0</b>	4.6	5.1	4.9	4.9	5.5	3.8	5.8	<b>4.9</b>	<b>4.9</b>	4.6	5.2	4.7	4.8	4.8	3.3	5.8
	0.1	5.2	5.2	4.6	<b>5.0</b>	5.2	4.9	5.4	3.6	5.7	<b>5.0</b>	5.2	5.2	5.4	5.2	4.9	5.2	3.4	6.4
	0.5	5.4	5.3	4.8	<b>5.1</b>	4.8	5.3	5.2	4.1	6.1	4.8	<b>4.9</b>	4.7	<b>4.9</b>	4.8	4.8	4.6	3	5.9
	0.9	5.3	5.3	5.3	4.9	2.9	<b>5.0</b>	5.1	4.3	5.3	<b>5.1</b>	4.8	4.8	<b>5.1</b>	1.8	4.6	<b>4.9</b>	3.2	6.2
Laplace	-0.9	5.2	4.8	5.3	6.4	<b>5.0</b>	4.7	5.6	4	6.6	5.3	<b>5.0</b>	5.8	<b>7.4</b>	5.1	5.2	<b>5.0</b>	3.8	<b>7.8</b>
	-0.5	<b>5.1</b>	4.6	4.8	6.1	4.8	4.7	5.4	3.8	6.5	<b>4.9</b>	4.4	<b>5.1</b>	6.3	4.8	<b>4.9</b>	4.8	3.3	<b>7.2</b>
	-0.1	4.9	4.1	4.4	4.9	4.9	4.7	<b>5.0</b>	3.5	6	<b>4.9</b>	4.1	4.7	5.5	4.6	4.5	4.7	3.1	6.6
	0.1	5.3	<b>4.9</b>	4.6	<b>5.1</b>	<b>4.9</b>	4.8	5.6	3.8	6.2	4.7	4.3	4.3	5.2	4.7	4.9	<b>5.0</b>	3.2	6.2
	0.5	5.3	4.8	4.5	<b>5.0</b>	4.8	5.2	5.6	4.3	5.9	<b>5.0</b>	4.5	4.4	4.6	4.2	4.5	4.8	3	5.8
	0.9	<b>5.0</b>	4.3	4.1	3.9	3.6	4.9	4.7	3.8	4.8	<b>5.1</b>	4.4	4.6	<b>5.1</b>	2.7	4.8	4.7	3.3	6.2
Exponential	-0.9	4.8	5.6	<b>5.0</b>	<b>7.4</b>	4.9	4.6	5.2	3.5	<b>6.9</b>	5.2	<b>4.9</b>	5.5	<b>8</b>	<b>5.1</b>	<b>4.9</b>	<b>8.3</b>	2.9	<b>7.4</b>
	-0.5	5.2	<b>8.5</b>	<b>5.1</b>	<b>7.9</b>	5.2	4.6	5.4	3.8	<b>7</b>	5.5	<b>8.3</b>	5.7	<b>8</b>	5.7	<b>5.4</b>	<b>7.8</b>	3.2	<b>7.1</b>
	-0.1	<b>4.9</b>	<b>8.2</b>	4.8	<b>7.2</b>	<b>5.1</b>	4.3	5.3	3.9	<b>6.9</b>	<b>5.0</b>	<b>7.8</b>	5.1	<b>6.8</b>	5.6	5.3	<b>9</b>	3.1	6.7
	0.1	<b>5.0</b>	<b>7.5</b>	4.9	<b>7.2</b>	4.8	4.6	5.5	3.6	6.7	5.7	<b>6.8</b>	<b>4.6</b>	6.3	5.9	<b>5.4</b>	<b>8.4</b>	3.2	6.2
	0.5	5.2	<b>6.9</b>	5.9	6.4	<b>5.0</b>	4.9	5.6	4.2	6.5	5.8	6.5	<b>4.4</b>	6.2	5.8	5.7	<b>8.9</b>	3.7	6.4
	0.9	<b>4.7</b>	5.7	<b>6.9</b>	5.7	4.4	4.3	<b>5.3</b>	<b>4.7</b>	5.8	<b>7.1</b>	<b>7.2</b>	<b>5.1</b>	<b>7.4</b>	4.7	<b>6.8</b>	<b>10.3</b>	4.5	<b>7.3</b>
Chisquare	-0.9	4.5	5.2	<b>5.0</b>	5.7	4.8	4.7	4.8	3.5	6.3	5.3	5.5	5.3	6.6	<b>4.9</b>	<b>4.9</b>	<b>5.1</b>	3.5	<b>7.4</b>
	-0.5	<b>5.1</b>	5.7	<b>4.9</b>	5.5	<b>5.1</b>	<b>5.1</b>	5.3	3.8	6.2	<b>5.1</b>	5.4	5.3	6	5.2	5.3	4.7	3	<b>6.8</b>
	-0.1	5.3	5.5	<b>5.0</b>	5.4	5.2	5.1	5.2	3.8	6.1	5.7	5.5	5.4	5.6	5.3	5.1	<b>5.0</b>	3.1	6.6
	0.1	<b>5.1</b>	5.3	4.8	5.2	4.8	<b>4.9</b>	5.2	3.5	5.7	5.5	5.3	<b>5.1</b>	5.6	5.3	<b>5.1</b>	<b>5.1</b>	3.1	6.7
	0.5	4.7	4.7	4.5	4.7	4.5	4.7	<b>5.1</b>	3.6	5.3	5.6	5.4	4.9	5.4	4.9	5.1	<b>5.0</b>	3	6.4
	0.9	5.5	5.3	5.2	<b>4.9</b>	3.2	<b>5.1</b>	5.2	4.4	5.4	5.4	4.9	4.6	<b>5.0</b>	2	5.2	4.7	2.7	6

**Table 6** Power simulation results ( $\alpha = 0.05$ ) of the tests for different distributions under varying correlation values ( $\rho$ ) with sample sizes  $(n_c, n_u) = (10, 10)$  and homoscedastic covariance matrix  $\Sigma_1$  under the MCAR framework. Values of too liberal tests corresponding to **red** values in the  $\Sigma_1$ -column from Table 1 in paper are printed in **red** colour.

Dist	$\rho$	$\delta = 0.5$								$\delta = 1$									
		$F$	Parametric bootstrap				Alternatives				$F$	Parametric bootstrap				Alternatives			
			$T_W^*$	$T_A^*$	$T_M^*$	$T_L^*$	$T_t$	$T_N$	$T_F$	$T_P$		$T_W^*$	$T_A^*$	$T_M^*$	$T_L^*$	$T_t$	$T_N$	$T_F$	$T_P$
Normal	-0.9	19.6	13.4	15	16.5	18	10.9	17.6	9.4	<b>21.6</b>	59.2	41	45.5	50	55.6	30.3	48.5	34.7	<b>58.6</b>
	-0.5	24.2	14.2	18.8	19.2	19.5	13.5	<b>20.7</b>	11.8	<b>24</b>	69.2	44.5	54	55.4	56.8	37.8	<b>54.8</b>	41.5	<b>62.9</b>
	-0.1	30.7	15.8	21.2	20.4	20.2	16.3	22	13.9	25.1	81.1	50.8	62.9	61.7	61.7	48.4	61.6	49.5	68.5
	0.1	35.2	17.2	23.1	22.1	21.8	18.7	24.1	15.1	26.6	88.2	55.9	68.7	66.5	65.7	55.6	66.6	55	73
	0.5	56.7	25.2	27.6	29.8	30.5	29	32.8	22.9	33.9	98.9	76.2	80.6	82.7	82.7	80.2	81.6	74	86.3
	0.9	99.7	81.2	31.5	81	70	88.6	71.5	72.5	81	100	100	94	100	99.9	100	99.4	99.9	100
Laplace	-0.9	20.8	13.9	16.2	18.4	19.2	12	21.7	11.5	<b>23.5</b>	60.9	46.2	49.2	54.6	57.5	34.4	58.1	41.9	<b>62.1</b>
	-0.5	25.1	15	19.4	20.3	20.7	14.1	24.8	14.4	<b>25.6</b>	69.9	48.6	56.5	59.2	59.2	40.4	64.2	48.4	<b>66</b>
	-0.1	32.4	17.5	23.1	23	22.4	17.6	30.3	18.6	28.5	82.1	55.7	65.6	65.6	64.1	51.3	72.2	57.8	72.4
	0.1	37.1	19.2	25	24.7	24.2	20.3	32.3	20.9	30.4	87.8	60.2	69.7	69.8	67.9	58.6	76	62.6	76
	0.5	58.3	30.3	30.4	34.5	34	32.9	42.5	30.7	39.2	98.4	79.6	81	83.8	83.6	80.7	87.3	78.8	86.9
	0.9	99.3	83.2	33.3	81.4	73.6	87.7	76.8	75.1	82.1	100	99.9	91.6	99.8	99.7	100	99.3	99.4	99.9
Exponential	-0.9	22.1	13.6	18.4	20.7	21.9	12.8	26.8	13.7	<b>25.1</b>	61.3	45.8	51.2	55.3	59.6	36.5	62.8	43.2	<b>62.4</b>
	-0.5	27	15.8	24	23.2	25.7	15.6	31	16.4	<b>28.6</b>	71.1	48.3	59	59.6	62	44.4	68	49.8	<b>66.1</b>
	-0.1	33.8	20	28.7	25.7	29.4	19.9	36.2	20.6	<b>31.5</b>	82.5	53.8	66.4	64.9	64.8	54.5	73.5	55.7	<b>71</b>
	0.1	39.6	23.1	30.8	27.8	31.7	23.2	39.2	22.7	33.7	87.4	57.9	69.6	67.7	66.6	60.5	76.7	59.5	73.1
	0.5	59.7	32.8	35.8	36.4	38.3	35.5	46.3	29.8	40.8	97.9	74.4	78.8	79.9	77.7	81.4	85	71.9	83.9
	0.9	99	80.7	37.2	78.3	<b>72.2</b>	86.9	73.4	67	78.7	100	99.9	89.6	99.4	<b>99.1</b>	99.9	97.9	97.8	99.3
Chisquare	-0.9	20.8	13.4	16.5	18	19.8	12.1	<b>18.8</b>	10.4	<b>23.2</b>	60.2	42.2	46.7	51.6	57	31.4	<b>49.7</b>	36.2	<b>59.6</b>
	-0.5	24	13.8	19.7	19	20.4	13.4	20.5	11.6	<b>23.9</b>	69.7	43.7	54.2	54.9	57.4	38	54.8	40.9	<b>62.6</b>
	-0.1	30.4	16	22.7	20.7	22	16.1	22.4	13.7	25.4	82.1	50.3	63.6	61.8	62.2	48.2	61.5	48.2	69
	0.1	35.9	17.8	24.6	22.5	24.1	18.6	24.5	15.7	27.5	88.6	53.9	67.5	65	65.2	54.7	65.2	52.3	71.6
	0.5	56.6	26.5	29.1	30.4	31.6	28.9	32.4	23.2	34.6	98.8	75.3	79.6	81.5	81.1	80.5	80.7	71	85.3
	0.9	99.6	80.1	32.4	79.8	70.3	87.6	69.7	68.3	80	100	100	92.9	99.9	99.7	100	99	99.6	99.9



**Table 7** Power simulation results ( $\alpha = 0.05$ ) of the tests for different distributions under varying correlation values ( $\rho$ ) with sample sizes  $(n_c, n_u) = (10, 30)$  and homoscedastic covariance matrix  $\Sigma_1$  under the MCAR framework. Values of too liberal tests corresponding to **red** values in the  $\Sigma_1$ -column from Table 1 are printed in **red** colour.

Dist	$\rho$	$\delta = 0.5$								$\delta = 1$									
		$F$	Parametric bootstrap				Alternatives				$F$	Parametric bootstrap				Alternatives			
			$T_W^*$	$T_A^*$	$T_M^*$	$T_L^*$	$T_t$	$T_N$	$T_F$	$T_P$		$T_W^*$	$T_A^*$	$T_M^*$	$T_L^*$	$T_t$	$T_N$	$T_F$	$T_P$
Normal	-0.9	34.9	21.2	15.6	<b>21.9</b>	29.5	11.3	<b>19.3</b>	12.5	<b>26.3</b>	88.4	69.6	45	<b>62.2</b>	82.3	30.5	<b>51.5</b>	46.2	<b>67.3</b>
	-0.5	43.8	18.7	18.9	23.8	25.9	13.2	21.3	14.6	<b>28.5</b>	94.5	61.1	56.2	66.2	72.3	37.6	59	53.2	<b>72.8</b>
	-0.1	54.7	18.6	22.7	24.6	24	15.9	<b>23.9</b>	16.7	<b>29.8</b>	98.4	59.5	66.4	70.1	69.7	47.5	<b>65.5</b>	59.5	<b>76.9</b>
	0.1	63.2	20.3	26.7	27.2	26.3	19.1	<b>26.9</b>	19.2	<b>32.2</b>	99.5	63.6	73.9	74.7	72.4	55.8	<b>71.8</b>	65.2	<b>81</b>
	0.5	87.1	27.8	34.6	34.9	33.9	29.7	36.6	26	40	100	79	87.6	86.2	85.5	79.9	85.9	79.7	91
	0.9	100	80.7	47.3	79.4	79.8	88.3	77.6	67.2	83.3	100	100	98.6	99.9	100	100	99.9	99.8	100
Laplace	-0.9	36.5	23.1	16.4	<b>23.7</b>	32.3	11.6	<b>23.6</b>	15.2	<b>28.9</b>	89	72	49	<b>65</b>	83.3	33.4	<b>61</b>	54.3	<b>69.9</b>
	-0.5	42.7	19.7	19.8	25	26.5	13.5	27.2	18.5	<b>30.3</b>	93.9	62.9	57.2	67.2	72	39.8	67.4	61.2	<b>73.6</b>
	-0.1	55.7	21.3	25.4	28.3	27.8	17.3	31.8	23.5	<b>34</b>	98.1	64.1	68.7	73.2	71.6	51.1	75.4	69.1	<b>79.3</b>
	0.1	64.4	22.6	28.6	30.5	28.6	20.1	35.6	26	<b>36.2</b>	99.4	67	75.1	77.2	74.3	58.8	80.8	73.9	<b>82.7</b>
	0.5	86.4	32.4	38.5	39.3	38.5	32.1	46.2	34.9	44.4	100	81.5	87.5	86.8	85.9	80.9	90.7	84.1	90.9
	0.9	100	83.7	49.2	79.3	80.7	87.6	82.8	71.3	83.6	100	99.9	96.8	99.4	99.9	100	99.7	99.2	99.9
Exponential	-0.9	37.9	25.6	19.9	28.2	37.5	13.3	<b>29.1</b>	18.5	<b>32.6</b>	88.5	73.6	52.3	65.8	84.8	36.6	<b>66</b>	57.3	<b>70.5</b>
	-0.5	45.6	26.9	25.9	<b>32.1</b>	37.2	15.8	33.3	23	<b>35.1</b>	94.1	65	61.3	<b>68.6</b>	75.6	44.5	70.9	62.5	<b>72.8</b>
	-0.1	56.5	<b>28.5</b>	32.2	<b>34.3</b>	<b>35.8</b>	19.3	38.3	26.9	<b>37</b>	97.8	<b>62.5</b>	68.3	<b>70.4</b>	<b>69.9</b>	54.3	76	66.9	<b>74.7</b>
	0.1	65.2	<b>30.6</b>	36.3	<b>36.9</b>	<b>37.5</b>	22.6	41.4	29.3	<b>39.2</b>	99.2	<b>65.1</b>	73.5	<b>74.4</b>	<b>71.8</b>	62.1	79.6	71	<b>78.4</b>
	0.5	86.3	<b>36</b>	44.1	43.5	<b>41.4</b>	34.3	49.9	34	46	100	<b>74.5</b>	82	81.9	<b>78.4</b>	80.5	87.4	78.1	85.3
	0.9	100	<b>80.3</b>	53	78.2	<b>76.3</b>	87.2	78	63.8	80	100	<b>99.9</b>	92.5	97.9	<b>99.9</b>	99.8	98.7	97.3	99.4
Chisquare	-0.9	36.1	23	16.4	<b>23.3</b>	32.5	11.9	<b>19.8</b>	13.4	<b>27.8</b>	89	70.5	47.2	<b>63.4</b>	83.5	31.8	<b>52.4</b>	48	<b>68.5</b>
	-0.5	43.2	19.7	20.2	24.5	28	13.1	<b>21.5</b>	14.5	<b>29.2</b>	94.6	60.3	55.9	65.5	71.7	37.8	<b>58.3</b>	52.3	<b>71.3</b>
	-0.1	54.3	20.3	24.3	25.8	26.4	16.2	24.6	16.8	<b>30.8</b>	98.4	60	66.9	69.6	69.5	48.1	65.5	58.7	<b>76</b>
	0.1	64.5	21.8	28.5	28.6	28.3	18.4	26.2	19.1	33.2	99.7	62.7	72.7	73.3	71.4	55.7	70.7	62.5	79.1
	0.5	87.4	29	36.2	35	35.3	29.9	35.6	25.8	40.2	100	77.2	86.1	84.9	83.2	80.4	84.2	76.9	89.4
	0.9	100	79.9	47.8	78.6	78.3	88.2	75.2	63	81.9	100	100	97.2	99.8	100	100	99.6	99.2	100

**Table 8** Power simulation results ( $\alpha = 0.05$ ) of the tests for different distributions under varying correlation values ( $\rho$ ) with sample sizes  $(n_c, n_u) = (30, 10)$  and homoscedastic covariance matrix  $\Sigma_1$  under the MCAR framework. Values of too liberal tests corresponding to **red** values in the  $\Sigma_1$ -column from Table 2 are printed in **red** colour.

Dist	$\rho$	$\delta = 0.5$								$\delta = 1$									
		$F$	Parametric bootstrap				Alternatives				$F$	Parametric bootstrap				Alternatives			
			$T_W^*$	$T_A^*$	$T_M^*$	$T_L^*$	$T_t$	$T_N$	$T_F$	$T_P$		$T_W^*$	$T_A^*$	$T_M^*$	$T_L^*$	$T_t$	$T_N$	$T_F$	$T_P$
Normal	-0.9	35.9	24.5	34.2	31.8	35.3	27.9	31.5	17.8	38.1	88.9	76.8	86.8	85	88	78.3	82.9	71.6	89.4
	-0.5	42.7	27.7	36.5	35.5	39.8	32.8	35.2	22	41.4	94.5	84.5	90.7	90.2	93	86.7	88	80.5	93.3
	-0.1	54.6	35.4	40.1	42.6	48.9	42.6	40.4	29.7	48.1	98.6	92.6	94.3	95.5	97.1	94.6	93	90.5	96.7
	0.1	63.6	42.6	43.1	49.4	55.9	50.9	45.8	37.1	54.2	99.6	96.2	95.8	97.6	98.7	97.4	95.4	95.2	98.3
	0.5	87	65.7	45.7	69.9	78.3	75.3	60.7	61.9	73.1	100	99.8	98.2	99.9	100	99.9	99.2	99.8	99.9
	0.9	100	100	47	100	99.9	100	96.1	99.9	100	100	100	99.4	100	100	100	100	100	100
Laplace	-0.9	36.2	24.6	33.5	33	34.8	27	39.2	21.4	38.9	88.3	77.8	86.5	85.5	87.3	77.5	90.6	77.3	89.6
	-0.5	43.6	29.8	37.9	38.1	41	33.6	45.5	28.2	43.7	94.1	84.9	89.9	90.2	91.7	85.5	93.7	85.9	92.8
	-0.1	55.2	38.1	42.1	45.7	50	44.1	54.8	38.8	50.7	98.2	92.7	93	95.3	96.5	94.2	97.3	94.3	96.6
	0.1	64.8	45.8	45.4	52.3	57.8	52.4	60.3	47.6	56.9	99.3	96	94.6	97.6	98.3	96.9	98.3	96.9	98.2
	0.5	86.6	67.6	48.3	71.7	78.7	75.6	73.8	70.7	74.8	100	99.8	97.1	99.8	99.9	99.9	99.7	99.9	99.9
	0.9	100	100	48.8	100	99.8	100	97.7	99.9	100	100	100	98.6	100	100	100	100	100	100
Exponential	-0.9	37.9	24.2	35.5	34.4	37.6	29.7	50.5	25.9	39.3	88.3	76.7	86.6	85.4	87.5	77.8	94.1	80.5	89.5
	-0.5	44.8	27.1	38.3	36.4	42.7	35.7	57	31.6	40.6	93.7	81.7	90.6	89.6	91.2	86	96.8	87.1	92
	-0.1	57	35.5	43	43.8	52.1	45.7	66.2	40.7	47.9	97.9	89.5	94.2	94	95.4	93.5	98.3	92.9	95.5
	0.1	65.2	42.6	45	49	57.9	52.8	70.4	47.9	53.5	99.2	93	95.6	95.8	97	96.2	98.9	95.7	97.2
	0.5	86.4	65.8	49.2	69.2	76.6	76.2	80.5	69.2	72.3	100	99.2	98	99.5	99.7	99.7	99.8	99.5	99.6
	0.9	100	99.8	50.3	99.7	99.2	99.9	98.1	99.6	99.7	100	100	99.6	100	100	100	100	100	100
Chisquare	-0.9	36.5	23.4	33.7	31.1	35.5	28.1	32.4	17.8	37.5	88.1	75.6	86.4	84.2	87.5	77.3	83.5	71	88.8
	-0.5	43.3	27.5	37.4	35.6	41	34.1	36	22.4	41.3	94.4	83.4	91.1	90.1	92.7	86.6	88.4	79.8	93.2
	-0.1	55.7	36	41.3	43.1	50.3	43.9	42.7	30.5	48.5	98.7	92.1	94.9	95.3	97	94.7	93.2	89.7	96.7
	0.1	64.2	41.8	43.1	48.7	56.5	51	46.8	37.2	53.5	99.6	95.8	96.2	97.5	98.6	97.6	95.5	94.3	98.2
	0.5	87	65.1	45.1	69.3	77.4	75.5	60	60.4	72.4	100	99.9	98.7	100	100	100	99.2	99.7	100
	0.9	100	100	46.5	100	99.8	100	96.1	99.9	99.9	100	100	99.7	100	100	100	100	100	100

**Table 9** Power simulation results ( $\alpha = 0.05$ ) of the tests for different distributions under varying correlation values ( $\rho$ ) with sample sizes  $(n_c, n_u) = (10, 10)$  and heteroscedastic covariance matrix  $\Sigma_2$  under the MCAR framework. Values of too liberal tests corresponding to **red** values in the  $\Sigma_2$ -column from Table 1 in paper are printed in **red** colour.

Dist	$\rho$	$\delta = 0.5$								$\delta = 1$									
		$F$	Parametric bootstrap				Alternatives				$F$	Parametric bootstrap				Alternatives			
			$T_W^*$	$T_A^*$	$T_M^*$	$T_L^*$	$T_t$	$T_N$	$T_F$	$T_P$		$T_W^*$	$T_A^*$	$T_M^*$	$T_L^*$	$T_t$	$T_N$	$T_F$	$T_P$
Normal	-0.9	15	10.5	11.2	12.6	13.4	9	14.2	7	17.5	44.1	29.1	31.2	34.9	39.8	22	34.4	21.6	44.2
	-0.5	18.2	10.9	13.9	14.6	13.8	10.8	15.4	8.3	19.3	53	29.8	38.1	39.7	39.8	27.8	39	26.2	47.9
	-0.1	22.3	11.4	15.2	15.3	14.1	12.4	16.9	9.2	19.9	64.5	32.6	44	43.6	41.8	34.8	43.9	30.4	51.8
	0.1	25	12.1	16.1	15.6	14.9	13.8	17.2	9.7	20.4	73	36.2	47.5	46.8	44.5	40.1	47.4	33.6	55.5
	0.5	39.7	16.2	19.2	20.2	19.4	20.4	22.3	13.4	25.3	92	51.3	57.1	60.1	58.8	61	59.5	47.5	67.2
	0.9	88.7	48.9	21.8	44.8	34.7	56	39	30.9	47.9	100	97.4	69.2	95.4	93	98.7	87.2	86.6	95.5
Laplace	-0.9	15.8	10.8	12	13.8	14	9.7	16.7	8.1	19	46.5	33.3	35.1	39.4	43.2	25.3	44	27.8	48.4
	-0.5	19	11	14.4	15.5	14.7	10.9	19	10.2	20.5	54.6	33.5	40.8	43	42.4	29.7	49.3	32.9	51.4
	-0.1	23.4	12.7	16.8	16.8	15.6	13.7	22.6	12.7	22.6	66.6	38.2	48.1	48.6	46	37.9	55.6	40	56.7
	0.1	26.7	13.2	17.6	17.8	16.6	14.8	23.1	13.4	23.8	73.5	41.4	51.4	51.5	49.1	43.5	59.6	44.1	59.2
	0.5	42.1	19.3	21.3	23.6	22	23.5	30	19.2	29.3	91.7	58	61.7	64.7	62.8	64.4	70.7	56.6	70.9
	0.9	87.5	54.1	22.7	50.3	42.4	60.1	49.7	39.2	53.1	99.9	96.8	70.5	93.3	92.1	97	89.7	85.3	93.9
Exponential	-0.9	19.6	13.9	17.1	19.1	19.3	13.2	27.2	13.1	23.2	48	37	39.4	43.3	47	30.5	53.6	33.1	50.4
	-0.5	23.1	15	21.4	21.4	21.9	15.2	30.6	15	25.9	57.1	37.1	46.8	47.9	49.2	36.2	57.9	37.8	54
	-0.1	27.8	17.9	25.3	23.5	24.4	18.8	34.1	18.1	28.2	67.1	41.2	52.2	51.4	50.3	43.8	62.1	42.4	57.4
	0.1	31.5	19.9	26.7	25.1	25.9	21.5	36.1	19.6	29.7	72.7	44	54.8	53.3	51.8	48.3	64.7	44.3	59.4
	0.5	45.3	26.4	30.5	30.8	29.6	30.4	41.2	23.7	34.9	87.5	57	61.7	63.1	60.1	64.5	72.3	53.8	68.3
	0.9	81.2	57.6	29.5	52.4	45.9	60.4	54.4	39.1	54.2	99.6	94.4	69.1	87.1	87	92.5	86.1	77.2	88.1
Chisquare	-0.9	16.5	11.3	13.7	14.9	15.3	10.7	16.2	8.4	19.8	46.2	31.7	34.2	37.8	42.4	24.6	37.7	24.7	45.7
	-0.5	18.6	11	15.7	15.8	15.6	11.4	17.2	9.2	20.5	53.4	30.6	39	40.4	41.2	29.1	41.9	27.3	48.4
	-0.1	22.6	12.3	17.5	16.6	16.3	13.5	18.4	10.1	21	65.4	33.9	45.6	44.4	43.4	36.1	46.2	31.3	52.5
	0.1	26.6	13.6	18.6	17.4	17.2	15	20.6	11.6	22.1	72.3	37	48.5	47.1	45.9	40.8	48.6	34.8	54.8
	0.5	40.6	18.3	21.3	22.5	21.4	21.9	24.9	15.2	26.9	90.9	51.9	58.2	60.3	58.9	61.1	60.7	46.9	66.5
	0.9	86.1	49.9	23.4	45.7	35.9	55.8	41.7	32.2	48.2	100	96.2	68.3	92.3	90.4	97	84.5	81.6	92.7

**Table 10** Power simulation results ( $\alpha = 0.05$ ) of the tests for different distributions under varying correlation values ( $\rho$ ) with sample sizes  $(n_c, n_u) = (10, 30)$  and heteroscedastic covariance matrix  $\Sigma_2$  under the MCAR framework. Values of too liberal tests corresponding to **red** values in the  $\Sigma_2$ -column from Table 1 are printed in **red** colour.

Dist	$\rho$	$\delta = 0.5$								$\delta = 1$									
		$F$	Parametric bootstrap				Alternatives				$F$	Parametric bootstrap				Alternatives			
			$T_W^*$	$T_A^*$	$T_M^*$	$T_L^*$	$T_t$	$T_N$	$T_F$	$T_P$		$T_W^*$	$T_A^*$	$T_M^*$	$T_L^*$	$T_t$	$T_N$	$T_F$	$T_P$
Normal	-0.9	25.2	14.9	11.7	15.9	20.5	9.4	15.7	9.8	20.5	74.5	50.2	31.2	41.2	64.5	22.4	36.8	30.2	48.5
	-0.5	32.2	13.1	13.7	17	16.8	10.5	16.4	10.3	21.7	84.3	39.3	38.3	45.3	49.7	27.8	42.4	34.5	53.4
	-0.1	40.4	12.4	16	17.3	15.8	12.4	18.2	11.8	22.2	92.3	36.2	44.9	47.7	44.8	34.6	46.4	37.8	56.2
	0.1	46.9	13.7	17.8	18.9	16.6	14.1	19.7	13	23.7	96.1	39.9	50.6	51.5	47.7	40.8	51.7	42.5	61
	0.5	69.1	17	20.9	21.9	20.6	20.5	24.2	15.8	27.4	99.8	51.9	62.1	62.6	59.8	61.5	63.8	53	71.6
	0.9	99.4	50.1	26.2	44.3	49.6	56	43	30	49.6	100	97.6	76.7	92.3	98.2	98.7	89.9	83.4	95.5
Laplace	-0.9	27.3	16.6	12.4	17.3	23	9.5	18.8	11.2	22.1	75.3	52.9	34.9	46.1	65.7	24.9	46	37.1	53.1
	-0.5	31.3	13.3	14.1	17.7	18	10.5	20.8	13.1	23.1	83.2	42.6	40.6	48.1	51.8	29.5	51.1	42.5	56.1
	-0.1	41.5	14	17.7	20	17.9	13.6	24.3	16.5	25.2	92.5	41.6	49.1	53.3	49.7	37.8	58.7	49.3	61.2
	0.1	48.5	14.7	19.3	21	18.3	14.8	26.1	17.5	26.2	96	44.2	55.2	56.5	52.9	44.3	63.7	53.1	64.7
	0.5	69.9	20	24.4	26.2	24.2	22.7	33	22.3	31.7	99.7	57.9	66.5	67.3	64.9	65	75.4	62.5	74.5
	0.9	99.4	56.6	28.6	49.6	53.4	60	53.4	36.9	54.5	100	97.2	77.1	89.4	97.8	96.8	91.6	84	93.7
Exponential	-0.9	29.2	22.1	18.1	23.6	30.4	13.8	29.1	18.1	28.2	75	58.2	41.1	50	70.1	30.8	55.8	44.2	55.8
	-0.5	35.1	22.5	22.7	26.6	29.5	15.7	33.2	21.7	30.3	82.2	49.5	47.3	53.4	60.3	36.1	60.1	49.4	59
	-0.1	43.1	23.2	26.8	28.3	28.3	18.6	36.2	24.6	31.4	89.5	46.8	53.5	55.7	54.2	43.1	64.6	52.8	60.6
	0.1	49.8	24.5	29.9	30.5	28.9	21.3	38.4	26.1	33.6	93.6	48.9	57.8	58.9	55.3	49.4	68.1	55.8	63.9
	0.5	68	27.2	33.5	34.5	30.3	29.1	43.9	28.3	37.3	99	56.1	64	65.1	60.1	63.9	74.9	61.2	70.1
	0.9	97	56.6	37.8	54	51.7	60.4	58.2	40.6	56.1	100	96.5	73.5	84.6	95.7	92.5	88.8	79	88.5
Chisquare	-0.9	27.3	17.5	13.6	18	23.8	10.7	17.8	11.3	22.5	74.5	51.7	33.8	44.4	66	24.4	40.1	33.1	50.7
	-0.5	32.2	14.3	15.4	18.4	19.3	11.1	18.3	11.9	22.7	83	40.8	39.7	46.1	51.4	28.6	44.2	36.1	53.6
	-0.1	39.8	14.2	17.8	19	18.1	13.5	20.5	13.1	23.8	92	38.9	46.9	49.5	47.1	35.7	49.7	39.9	57.5
	0.1	47.4	15.4	19.9	20.5	19	14.8	21.8	14.5	25.7	96.2	40.5	51.3	52.2	48.6	41.6	53.7	42.8	59.8
	0.5	69.3	18.9	23.5	24.4	22.5	22.5	26.8	17.5	29.3	99.7	51.9	61.3	61.5	59	60.9	64.6	52	69.9
	0.9	99.3	50.8	28.6	46.1	49.9	56.6	45.6	31.4	50.6	100	97.1	74.4	89.6	97.3	97.1	87.5	79.6	93.1

**Table 11** Power simulation results ( $\alpha = 0.05$ ) of the tests for different distributions under varying correlation values ( $\rho$ ) with sample sizes  $(n_c, n_u) = (30, 10)$  and heteroscedastic covariance matrix  $\Sigma_2$  under the MCAR framework. Values of too liberal tests corresponding to **red** values in the  $\Sigma_2$ -column from Table 2 are printed in **red** colour.

Dist	$\rho$	$\delta = 0.5$								$\delta = 1$									
		$F$	Parametric bootstrap				Alternatives				$F$	Parametric bootstrap				Alternatives			
			$T_W^*$	$T_A^*$	$T_M^*$	$T_L^*$	$T_t$	$T_N$	$T_F$	$T_P$		$T_W^*$	$T_A^*$	$T_M^*$	$T_L^*$	$T_t$	$T_N$	$T_F$	$T_P$
Normal	-0.9	26.2	17.2	25.4	23.6	25.2	20.5	21	9.5	29.9	75.2	58.7	73.1	69.9	73.6	61.8	65.1	46.7	77.7
	-0.5	31.3	19.2	27.2	26	28.2	24	23.1	11.9	31.6	83.9	66	78.4	76.6	79.4	71.3	70.6	55.5	82.5
	-0.1	40.1	24	31	30.7	34	30.6	27.2	15.7	36.4	92.7	76.1	84.1	83.8	86.8	82.8	77.3	67.8	88.3
	0.1	47.2	28.1	33.3	35	39.2	36.8	30.4	19.6	40.9	96.3	83.9	87.2	89	91.6	89.7	82.1	76.4	92
	0.5	69	43.5	35.7	48.6	56.5	55.8	39.1	32.8	53.8	99.9	96.8	92.2	97.6	98.7	98.7	92.2	93.9	98.2
	0.9	99.5	94.4	37.9	92.2	89.8	97.4	69.4	84.9	93.2	100	100	96.5	100	100	100	99.8	100	100
Laplace	-0.9	26.5	18	25.2	24.8	25.6	20	27.6	12.5	<b>30.2</b>	74.6	59.9	72.7	70.8	73	62	77	55.3	<b>77.9</b>
	-0.5	32.2	20.8	28.9	28.4	29.2	24.7	32.2	16.4	34.2	83.1	67.4	77.8	76.8	78.4	70.9	82.6	65.9	82.5
	-0.1	40.7	25.4	31.9	33.3	35.1	31.8	39.6	22.9	38.9	92.1	77.2	83.6	84.6	86.6	83.1	89.3	78.3	88.6
	0.1	49	30.6	35.4	38.4	41.2	38.4	43.8	27.9	44.1	95.6	83.9	86.6	88.8	90.6	88.6	92.2	85.1	91.8
	0.5	70.1	46	37.6	51.5	58.2	57.3	54.6	44.3	56.5	99.7	95.9	91	97	98.5	98.3	97.3	96.2	97.9
	0.9	99.1	93.9	39.2	91.8	89.8	96.5	84.2	90.9	92.6	100	100	95.4	100	100	100	99.9	100	100
Exponential	-0.9	30.1	20.5	29	28.9	29.8	24.7	<b>50</b>	22.6	<b>33.2</b>	74.3	60.5	71.9	71.6	72.9	62.8	<b>87.6</b>	64.9	<b>76.7</b>
	-0.5	34.2	21.6	30.9	29.9	32.7	28.4	<b>55.3</b>	26.2	33.5	82	63.9	77.3	75.9	77.9	71	<b>91.3</b>	70.6	80.3
	-0.1	43.5	27.3	34.9	34.4	39	35	<b>62.6</b>	32.5	38.7	90	73	82.3	81.5	83.9	81.1	<b>94</b>	79.4	85.4
	0.1	50.2	31.8	36.9	38.3	43.3	40.4	<b>66.3</b>	37.4	42.4	93.4	78.6	84.7	85	87	86.2	<b>95.2</b>	83.9	88
	0.5	68.5	48.3	39.5	51.8	58.2	58.6	<b>74.6</b>	52.1	55.4	99	92.7	89.5	93.7	95.7	96.1	<b>97.6</b>	94	95
	0.9	<b>97.3</b>	<b>90.8</b>	40.2	<b>86.7</b>	85.2	92.8	<b>89.2</b>	86.8	<b>87.5</b>	<b>100</b>	<b>100</b>	95.5	<b>99.9</b>	100	100	<b>99.8</b>	99.9	<b>99.9</b>
Chisquare	-0.9	27	17.4	25.7	24	26.2	21	25.1	11.2	<b>30.4</b>	74.1	57.9	72.2	68.9	72.5	61.7	68.8	48.5	<b>76.9</b>
	-0.5	32.4	20.1	29	26.9	30.3	25.7	28.5	14.2	33.1	83.3	64.8	78.6	76.1	79.4	71.9	74.6	57.6	81.9
	-0.1	40.9	25.1	32.2	31.9	35.8	32.4	33.2	18.8	37.4	92.2	75.4	84.2	83.8	86.5	83	81.2	70	87.6
	0.1	47.5	29.3	34.1	36.1	40.4	37.6	35.9	22.5	40.9	95.6	81.7	86.6	87.7	90.4	88.5	85	76.4	91.1
	0.5	68.7	44.1	35.7	48.8	56.2	55.7	45.1	34.8	53.3	99.8	95.8	91.8	96.8	98.1	98.2	93.2	92.9	97.7
	0.9	99.1	93.6	37.5	90.6	89.1	96.4	74.8	85.2	91.7	100	100	96.7	100	100	100	99.7	100	100

**Table 12** Power simulation results ( $\alpha = 0.05$ ) of the tests for different distributions under varying correlation values ( $\rho$ ) with sample sizes  $n = 10$  and homoscedastic covariance matrix  $\Sigma_1$  under the MAR framework. Values of too liberal tests corresponding to **red** values in the  $\Sigma_1$ -column from Table 3 are printed in **red** colour.

Dist	$\rho$	$\delta = 0.5$								$\delta = 1$									
		$F$	Parametric bootstrap				Alternatives				$F$	Parametric bootstrap				Alternatives			
			$T_W^*$	$T_A^*$	$T_M^*$	$T_L^*$	$T_t$	$T_N$	$T_F$	$T_P$		$T_W^*$	$T_A^*$	$T_M^*$	$T_L^*$	$T_t$	$T_N$	$T_F$	$T_P$
Normal	-0.9	10	6.8	9.3	8.6	9.1	8.3	<b>12.8</b>	6.4	11.6	26	12.3	22.5	17.3	23.9	19.4	<b>29.4</b>	17.2	27.9
	-0.5	11.8	7.5	10.2	9.4	9.6	9.3	<b>13.8</b>	7.2	13.1	30.8	14.8	24.2	21.1	24.6	22.1	<b>31</b>	18.9	29.5
	-0.1	14.7	8.4	10.5	10	10.8	11.8	<b>15</b>	8.4	14	42.1	19.2	28.4	25.7	29.5	30.5	<b>36</b>	24	35.6
	0.1	16.4	9.2	10.8	10.5	11.1	13	<b>15.7</b>	9	14.8	51	23.6	31.6	29.8	33.9	36.4	<b>41</b>	28.7	40.9
	0.5	28.6	12.3	10.8	13.2	13.7	19.9	19.2	11.8	18.4	77.2	39.2	37.6	44.8	47.9	58.3	54	42.8	57.5
	0.9	88.2	48.4	12.6	46.9	<b>53.4</b>	69.7	41.9	38.6	56.8	100	97.4	45.6	97.3	<b>96.2</b>	99.8	85	81.6	98.3
Laplace	-0.9	11.5	6.7	10.4	9.9	10.8	9.3	<b>16.3</b>	8.7	14.2	29.8	15.5	27.4	22.2	28.1	22.6	<b>36.4</b>	22.8	33.5
	-0.5	13.6	6.7	11.1	10.4	11.6	10.3	<b>17.8</b>	9.6	<b>16.1</b>	37.3	19.3	30.9	28.2	31	27.4	<b>41.7</b>	26.6	<b>39.3</b>
	-0.1	16.7	7.9	11.6	10.9	10.8	12	<b>20.1</b>	10.8	18	48.3	24.9	36.2	33.7	36	35.6	<b>48.3</b>	33.8	46.4
	0.1	19.5	9.2	12.6	12	12.3	14.3	<b>21</b>	12.5	19.5	55.9	29.4	38.9	37.5	39.2	41.4	<b>51.8</b>	37.5	50.8
	0.5	33.2	14.5	14	16.1	16.6	23.5	26.9	17.4	25.7	80.8	48.3	46.6	52.8	54.3	64	63.4	51.3	65.8
	0.9	89	55.3	14.2	52.8	<b>60.1</b>	71.5	48.2	42.8	62.3	100	97.1	52.6	96.8	<b>95.8</b>	99.3	86.9	83.5	98
Exponential	-0.9	11.7	4.7	10.7	9.3	11.8	9.4	<b>18.2</b>	9.4	14.2	32.9	14.4	30.4	24.3	32.4	25.2	<b>41.3</b>	25.2	35.6
	-0.5	12.9	5.5	11	7.8	13.3	10.3	<b>19.2</b>	10.1	<b>14.5</b>	39.2	17.2	34.4	27.9	36.1	29.9	<b>45.5</b>	28	<b>38.9</b>
	-0.1	17.2	8.1	11.3	8.4	15.4	13.1	<b>21.7</b>	12.3	16.5	50.1	24	39.9	33.2	40.2	38.5	<b>50.3</b>	33.8	45.1
	0.1	20.5	9.5	12.5	10.1	17.5	15.6	<b>24.2</b>	13.9	18.7	58.8	31.2	44.2	38.6	45.5	45.2	<b>55.3</b>	39.1	51.2
	0.5	34.6	16.4	12.7	16.1	23.5	25.3	28.4	19.6	25.5	79.9	49.3	51	52.8	56.5	65.4	64	50.6	64.2
	0.9	88.9	58.1	11.4	54	<b>63.7</b>	73	47.5	42	61.6	99.9	97.4	58.5	95.7	<b>96.5</b>	98.9	88.4	84.4	96.4
Chisquare	-0.9	9.7	6.4	8.9	7.9	9.5	7.9	<b>12.8</b>	6.5	<b>11.5</b>	25.6	11.8	22.9	16.8	24.6	19.7	<b>29.8</b>	17.5	<b>27.4</b>
	-0.5	11.6	7.2	9.6	8.9	10.5	9.7	<b>14</b>	7.5	12.5	32.5	15.7	26.1	21.7	27.5	24.2	<b>32.6</b>	20.2	31.3
	-0.1	14.2	7.6	10.4	9.2	11.1	11.5	<b>15.1</b>	8.2	13.5	42.8	19.3	29.5	25.6	31.2	30.8	<b>37.2</b>	25.3	35.6
	0.1	16.8	8.7	10.8	9.9	12	13.1	<b>15.8</b>	9.2	14.5	50.9	24.1	32.6	30.3	35.6	37.2	<b>41.2</b>	29.2	40.6
	0.5	27.8	12.2	10.3	12.6	14.9	19.6	19.4	12.2	18.4	78	40.5	38.4	45.3	49.2	59.1	53.1	41.5	56.7
	0.9	88.4	49.2	11.7	47.5	<b>55.9</b>	70.6	41.3	37.8	56.2	100	97.2	46.7	97	<b>96</b>	99.8	84.6	81.2	97.9

**Table 13** Power simulation results ( $\alpha = 0.05$ ) of the tests for different distributions under varying correlation values ( $\rho$ ) with sample sizes  $n = 20$  and homoscedastic covariance matrix  $\Sigma_1$  under the MAR framework. Values of too liberal tests corresponding to red values in the  $\Sigma_1$ -column from Table 4 are printed in red colour.

Dist	$\rho$	$\delta = 0.5$								$\delta = 1$									
		$F$	Parametric bootstrap				Alternatives				$F$	Parametric bootstrap				Alternatives			
			$T_W^*$	$T_A^*$	$T_M^*$	$T_L^*$	$T_t$	$T_N$	$T_F$	$T_P$		$T_W^*$	$T_A^*$	$T_M^*$	$T_L^*$	$T_t$	$T_N$	$T_F$	$T_P$
Normal	-0.9	19.5	12	18.5	17.8	18.7	16.8	17.7	9.9	22.3	58.7	35.9	55.4	49.4	56.4	49.6	51.8	38.4	61
	-0.5	23.2	14.1	20	19.4	21.5	20	19.3	12.2	23.7	69.9	44	60.2	56.6	64.6	58.3	57.7	46.7	66.2
	-0.1	29.8	17.2	21.2	22.3	25.9	24.2	22.1	15.3	26.5	82	55.6	65	66.1	74.3	70.9	64	57	73.4
	0.1	35.9	20	21.6	24.4	29.1	28.3	24.2	18	29.1	88.2	64.2	67.1	72.2	80.4	77.6	68.9	64.5	78.2
	0.5	56.4	31	21.4	34.3	41.9	43.7	30.4	29	38.9	99.1	87.8	72.9	90.6	95.1	95.5	83.1	87.4	93
	0.9	99.7	92.7	19.4	92.7	84.4	97.6	69.2	91.4	93.3	100	100	75.4	100	100	100	99	100	100
Laplace	-0.9	21.2	13.2	20.8	21.6	20.2	17.6	22.6	12.3	25.1	59.4	41	57.1	55.9	57.6	50.2	61.2	44.4	64.5
	-0.5	24.8	14.6	21.7	22.8	23	19.5	25.9	14.7	27	69.8	49.4	62.9	62.9	64.9	58.8	68.9	53.8	69.8
	-0.1	31.4	18.4	23.4	25.3	27	24.3	30.4	19.4	30.9	82	61.3	69.2	71.7	75.1	70.5	76.8	66	77.8
	0.1	36.2	20.8	23.6	26.7	29.6	28.1	33.4	22.6	32	88.1	67.7	71.3	76.3	80	76.7	79.6	71.6	81.3
	0.5	58.3	35	26	38.4	44.1	46	42.6	36.8	44.4	98.4	88	75.9	90.6	94	93.5	89.3	88.9	93
	0.9	99.3	91.7	23.9	90.7	85.4	95.9	75.7	90.1	91.8	100	100	78.8	100	100	100	98.9	100	100
Exponential	-0.9	22.2	11.7	22.1	23.3	21.8	19.7	28.7	14.6	25.7	62.2	41.9	60.5	59.9	60.9	53.8	68.9	48.6	66.2
	-0.5	26.5	14.2	23.7	21.9	26	23	31.9	16.9	25.9	70.8	46.7	66.1	65.1	67.6	62.8	73.2	54.5	69.8
	-0.1	33.3	19	25	23.5	31.1	29.1	35.9	21	28.6	81.8	57.8	72.1	71.9	74.7	73.6	79.6	63.5	75.6
	0.1	39.8	22.3	26.1	26	35.7	33.8	39.6	24.9	32	88.1	65.3	73.5	75	78.9	78.7	82.5	68.5	78.4
	0.5	59.3	35.3	25	36	45.6	48.8	45.4	35.4	41.6	97.7	84.5	78.3	87.4	90	92.7	88.8	83.4	89.3
	0.9	98.9	90.8	19.8	87.6	84.5	95.1	73.4	83.8	88.6	100	100	82.8	100	99.9	100	99.3	99.9	100
Chisquare	-0.9	19.7	11.5	19.2	17.8	19.1	17.3	18.4	9.9	22.1	59.1	35.8	55.9	51.2	57.1	50	53.2	38.8	61.2
	-0.5	23.4	13.3	20	18.9	22	20.6	20.2	12.3	23	69.6	43.3	61.6	57.5	65.6	59.4	58.5	45.8	66.8
	-0.1	30.5	16.7	20.5	21	26.7	25.5	22.2	15.7	25.5	81.7	54.2	64.9	64.9	73.2	70.4	63.7	55.1	72.1
	0.1	36.2	19.4	20.2	23.1	30	29.7	24	18.1	27.6	89	63.5	68.5	72.1	79.5	78.1	69.3	63.5	77.5
	0.5	56.1	31.8	20.5	33.7	42.8	45.1	30.1	29.4	38.5	98.9	86.4	73.6	89.6	93.8	94.9	82.1	84.7	91.8
	0.9	99.7	92.3	18.8	91.4	84.7	97.4	67.7	88.2	91.9	100	100	77	100	100	100	99.2	100	100

**Table 14** Power simulation results ( $\alpha = 0.05$ ) of the tests for different distributions under varying correlation values ( $\rho$ ) with sample sizes  $n = 30$  and homoscedastic covariance matrix  $\Sigma_1$  under the MAR framework. Values of too liberal tests corresponding to **red** values in the  $\Sigma_1$ -column from Table 5 are printed in **red** colour.

Dist	$\rho$	$\delta = 0.5$								$\delta = 1$									
		$F$	Parametric bootstrap				Alternatives				$F$	Parametric bootstrap				Alternatives			
			$T_W^*$	$T_A^*$	$T_M^*$	$T_L^*$	$T_t$	$T_N$	$T_F$	$T_P$		$T_W^*$	$T_A^*$	$T_M^*$	$T_L^*$	$T_t$	$T_N$	$T_F$	$T_P$
Normal	-0.9	28.5	18.9	27.8	27.5	27.3	24.2	25	13.8	<b>31.9</b>	77.7	60.6	75.5	72.8	76.2	67.8	69.8	55.8	<b>79.8</b>
	-0.5	33.7	21.2	29.7	29.5	31.2	27.4	27.4	16.6	33.8	86.3	68.6	80	79.2	82.5	77.6	74.8	64.8	84
	-0.1	43.5	27.1	32.2	34.3	38.1	35.8	31.1	22.4	38.2	94.8	80.4	85.3	87	90.1	87.4	82.5	77.6	90
	0.1	50.4	31.6	32.9	37.4	43	40.5	33.4	26.4	41.4	97.8	86.9	88.1	91.5	94.4	92.9	86.4	84.4	93.2
	0.5	75.1	49.9	34.8	53.6	62.2	61.1	44.9	45.1	56.7	100	98.4	92.7	98.9	99.5	99.5	95.5	97.6	99.1
	0.9	100	99.4	31.7	99.4	97.7	99.8	87.1	99.1	99.4	100	100	96.2	100	100	100	100	100	100
Laplace	-0.9	29.7	19.7	28.6	30.2	28.6	23.6	30.2	16.2	33.7	78.4	63.5	76.5	76.1	76.8	67.7	80.2	63.5	81.5
	-0.5	35	23.2	31.6	33.8	32.3	27.6	36.6	21	37	86.6	72	81.9	82.2	82.9	76.2	85.9	73.6	86
	-0.1	44.3	28.2	35	37.2	38.7	35.2	42.6	28.2	41.3	94	81.8	86.1	88.2	89.9	86.4	91.2	83.6	90.8
	0.1	51.8	32.2	36.2	40.4	44.4	40.2	47.4	33.3	44.9	96.9	87.6	88.3	91.4	93.2	91.1	93.3	89	93.3
	0.5	75	50.8	38.4	55.1	62	60.5	58	52.9	59.2	99.9	97.7	93.1	98.5	99.1	98.9	98.1	98	98.8
	0.9	100	99	37.6	98.8	97.1	99.6	91.8	98.8	98.9	100	100	95.8	100	100	100	100	100	100
Exponential	-0.9	31.5	19	30.3	<b>31.9</b>	30.6	26.3	39.6	20.1	<b>34.8</b>	77.2	62	75.5	<b>76.8</b>	76	67.9	85.3	65.4	<b>80.2</b>
	-0.5	36.1	<b>21.8</b>	32.3	<b>31.8</b>	34.6	31.3	44.3	23.8	<b>34.8</b>	85.7	<b>67.7</b>	81.1	<b>81.7</b>	81.8	77.3	89.4	72.1	<b>83.9</b>
	-0.1	46	<b>28</b>	35.3	<b>35.6</b>	41.9	39.1	51.3	29.8	<b>39.3</b>	93.7	<b>77.6</b>	86.5	<b>87.4</b>	88.3	87.3	93.2	82	<b>89</b>
	0.1	53.1	<b>32.6</b>	36.8	<b>39.1</b>	46.9	45.6	54.8	34.5	43	96.6	<b>84</b>	88.9	<b>90.4</b>	91.7	91.8	94.9	86.5	91.8
	0.5	75.4	<b>51.3</b>	37.8	53.2	62.7	64.5	63.2	50.6	56.3	99.7	<b>96</b>	92.4	96.8	98.2	98.6	97.9	95.5	97.2
	0.9	100	98.5	<b>33.1</b>	97.4	95.7	99.3	90.5	96.3	97.4	100	100	<b>96.8</b>	100	100	100	100	100	100
Chisquare	-0.9	27.8	17.7	26.9	26.3	26.8	23.3	25	13.7	30.8	77.3	59.5	75.4	73.2	76.1	67.5	70.4	56	79.2
	-0.5	34.2	20.6	29	28.4	31.4	28.5	26.8	16.6	32.9	86.4	67.2	80.3	78.9	82.2	77	75.5	63.7	83.8
	-0.1	43.5	25.4	30.7	32.1	37.4	36	30.9	21.3	36.2	95.1	79.6	86.7	87.2	90.5	88.8	83.5	76.4	90.3
	0.1	51.2	31.1	32.2	36.4	43	42.2	33.3	26.3	40.3	97.5	85.9	88.6	90.8	93.7	92.9	86.8	82.7	92.9
	0.5	75.9	49.5	32.5	52.9	63.2	63.5	43.7	43.5	56	99.9	97.9	93.3	98.6	99.3	99.3	95.2	96.5	98.8
	0.9	100	99.4	29.2	99.4	97.6	99.9	85.7	98.6	99.2	100	100	96.9	100	100	100	100	100	100



**Table 15** Power simulation results ( $\alpha = 0.05$ ) of the tests for different distributions under varying correlation values ( $\rho$ ) with sample sizes  $n = 10$  and heteroscedastic covariance matrix  $\Sigma_2$  under the MAR framework. Values of too liberal tests corresponding to red values in the  $\Sigma_2$ -column from Table 3 are printed in red colour.

Dist	$\rho$	$\delta = 0.5$								$\delta = 1$									
		$F$	Parametric bootstrap				Alternatives				$F$	Parametric bootstrap				Alternatives			
			$T_W^*$	$T_A^*$	$T_M^*$	$T_L^*$	$T_t$	$T_N$	$T_F$	$T_P$		$T_W^*$	$T_A^*$	$T_M^*$	$T_L^*$	$T_t$	$T_N$	$T_F$	$T_P$
Normal	-0.9	8.2	6.7	8.1	8.1	7.6	7.1	10.7	5.2	9.9	18.4	10	17	13.3	16.8	14.8	20.9	11.7	20.4
	-0.5	9.4	6.8	8.4	8.3	7.3	7.8	11	5.5	10.3	22.8	11.8	19.2	16.7	18.4	17.3	22.6	13.5	23.2
	-0.1	11.1	7	8.4	8.3	7.8	9	11.6	5.9	11.4	31.1	14.1	21.8	19.7	20.4	22.2	26.2	16	27.2
	0.1	12.7	7.5	9.2	8.8	8.4	10.2	12.5	6.8	12	36.6	15.5	23	21.6	22	25.5	27.6	17.9	29.2
	0.5	19.5	9.1	8.9	10.2	8.9	13.6	14.1	7.8	14	57.9	25	26.9	29.8	28.8	40.4	35.3	25.4	39.5
	0.9	57.8	24.9	10.3	22.8	29.3	37.9	22.9	17.4	29.6	99	77.4	34.7	73	71.2	91.3	63.3	57	80.3
Laplace	-0.9	9	5.9	8.6	8.9	8.2	7.5	13.2	6.6	11.9	21.6	11.6	19.7	16.9	19.8	16.7	27.2	15.7	25.1
	-0.5	10.8	5.9	9.6	9	8.5	8.4	14.6	7.3	13.7	27	13.4	23.3	21.7	22.1	20.1	31.8	19.4	29.9
	-0.1	12.2	6	9.3	8.8	7.8	9.3	14.8	7.8	14	35.8	17	27.6	26.3	24.3	26.1	35.6	23.2	35.5
	0.1	14.5	7.1	10	9.6	8.7	10.8	16	8.9	15.6	41.4	20.4	28.9	27.7	26.5	30.3	38.1	25.9	38.4
	0.5	23.3	9.8	11.2	11.7	10.6	16.1	19.1	11.9	19.1	63.3	31.3	33.9	37.6	35.6	47	46	34.6	49.3
	0.9	63.4	31.1	11.9	30.2	38.8	45.9	31.5	25.7	38.7	98	81	40.3	77.5	77.7	90.6	69.8	62.6	83.9
Exponential	-0.9	12.8	5.8	12.4	10.9	12.8	10.7	19.3	9.6	15.3	27.7	14.4	26.5	22.7	27.2	22.8	35.2	20.7	30.5
	-0.5	13.5	7.2	13.1	10.7	13.8	12.1	20.9	10.5	15.5	31.3	15.9	29.6	25.8	28.7	26	38.9	21.9	32.6
	-0.1	15.7	8.9	13.7	10.8	14.9	13.7	22.1	11.8	16.9	39.8	21.3	34.4	29.6	32.5	31.9	42.5	26.3	37.7
	0.1	19.3	11.2	14.9	12.7	16.8	16.4	22.9	13.1	19	46.2	26.4	37.7	33.7	35.6	37.6	45.6	29.5	41.6
	0.5	28.8	16.7	15.2	17.1	20.2	24	25.9	16.4	23.4	64.7	39.1	42.2	43.2	42.6	52.2	51.5	36.4	51
	0.9	65.5	43.6	12.8	37.6	46.4	51.3	34.8	27	42.4	95.8	81.7	47.3	74	81.1	85.7	66.2	57.3	77.4
Chisquare	-0.9	8.5	6.5	8.4	7.8	8.2	7.3	11.2	5.5	10.4	20.2	10.2	18.6	14.8	18.6	15.8	23.2	12.7	22
	-0.5	9.7	6.8	9.2	8.5	8.8	9.2	12.6	6.3	11.2	24.1	12.1	20.5	17.5	19.7	18.5	24.5	14.4	24.2
	-0.1	11.7	7.7	9.6	9.1	9.3	10.2	12.8	6.8	12.1	31.3	14.2	23.2	20.5	21.7	23.1	27.4	16.6	27.5
	0.1	14.9	8.1	9.5	9.2	9.8	11.4	14	7.7	12.9	37.3	16.7	24.9	22.8	23.8	27.1	29.9	18.9	30.9
	0.5	21.6	10.8	10.5	11.7	11.4	16.5	15.8	9.9	15.8	59.6	27.8	30	32.4	32.7	44	38.4	27.9	42.2
	0.9	60.2	29.7	10.3	26.6	34.2	42.9	25.5	19.9	33.1	98.1	77.4	36.7	72.5	73.7	88.8	62	55.3	78.2

**Table 16** Power simulation results ( $\alpha = 0.05$ ) of the tests for different distributions under varying correlation values ( $\rho$ ) with sample sizes  $n = 20$  and heteroscedastic covariance matrix  $\Sigma_2$  under the MAR framework. Values of too liberal tests corresponding to **red** values in the  $\Sigma_2$ -column from Table 4 are printed in **red** colour.

Dist	$\rho$	$\delta = 0.5$								$\delta = 1$									
		$F$	Parametric bootstrap				Alternatives				$F$	Parametric bootstrap				Alternatives			
			$T_W^*$	$T_A^*$	$T_M^*$	$T_L^*$	$T_t$	$T_N$	$T_F$	$T_P$		$T_W^*$	$T_A^*$	$T_M^*$	$T_L^*$	$T_t$	$T_N$	$T_F$	$T_P$
Normal	-0.9	14.8	9.7	14.7	14.3	14	13.1	13.2	6.5	<b>18</b>	44.5	24.9	42.5	37.5	41.7	37.3	36.9	23.6	<b>48.3</b>
	-0.5	17.8	11.3	15.9	16.1	16.4	15.2	14.3	8	19.3	52.7	30.3	45.6	43	47	43.9	40	28.1	51.6
	-0.1	22	12.9	16.7	17	18.2	18.2	15.2	9.4	20.3	64.8	37.7	49.5	49.4	54.5	53.9	45.2	35.3	57.1
	0.1	25.8	14.2	16.9	18.5	20.6	21	16.6	10.8	21.9	73.5	44	51.5	54.5	60.7	61.4	49.2	40.8	61.4
	0.5	39.3	20.1	16.6	22.9	27.5	29.6	19.4	14.9	27.1	92.3	65.3	55.4	70.4	79	80.9	59.5	59.4	75.2
	0.9	88.6	61.4	14.6	53.8	42.1	71.8	33.2	41.6	56.9	100	99.6	59.1	99.1	97.9	100	88.1	97.8	99.4
Laplace	-0.9	15.4	10.2	15.6	17	14.3	13.8	16.3	8	<b>19.8</b>	46.5	28.8	45.4	43.9	44.1	39.7	46.8	29.7	<b>52.4</b>
	-0.5	18.3	11.1	16.7	18.8	16.7	15.7	18.7	9.8	<b>21.4</b>	55	35.1	50	50.1	49.8	46	53.2	36.9	<b>57.4</b>
	-0.1	23.5	12.8	18.5	19.9	19.6	19.1	22.4	12.7	23.5	66.3	42.5	54.4	55.9	57.2	54.9	60.2	44.9	62.4
	0.1	27.2	14.7	18.9	20.9	21.7	21.5	23.9	14.7	<b>25</b>	73.9	48.8	56.6	60.3	62.8	61.9	63.6	50.5	<b>66.5</b>
	0.5	42.4	22.2	19.8	26.7	29.8	32.3	29.1	20.8	31.9	91.5	68.3	60.3	73.4	79	80.6	73.4	67.1	78.1
	0.9	87.4	62.9	17.4	57.8	49.2	72.1	48.6	50.7	61.1	100	99	63.2	98	96.8	99.5	91.7	95.9	98.4
Exponential	-0.9	19.9	11.8	20.1	<b>21.9</b>	19.4	17.7	<b>28.6</b>	13.2	<b>24</b>	48.5	33.3	47.5	<b>49</b>	46.9	42.5	<b>58.2</b>	35.2	<b>53.6</b>
	-0.5	22.7	13.5	21.9	21.4	23	21.4	<b>31.4</b>	14.9	24.1	56	34.8	52	52.5	52.2	49.2	<b>61.4</b>	39.5	56.1
	-0.1	27.6	<b>17.1</b>	23.1	22.9	26.2	25.4	<b>34.6</b>	17.7	25.8	66.8	<b>44.1</b>	56.8	57.5	59.4	58.8	<b>67.2</b>	47.2	61.2
	0.1	31.7	19.7	22.9	23.9	28.2	27.9	<b>35.9</b>	19.8	27	72.9	49.3	58.6	60.6	63.2	63.6	<b>69.5</b>	50.7	64
	0.5	45.2	29.4	22.4	30.3	35.5	38.7	<b>40.2</b>	25.9	33.6	88	67.5	61.7	70.8	75	78.3	<b>75.8</b>	63.2	73.4
	0.9	<b>82.1</b>	<b>66</b>	16.8	<b>54.6</b>	52.9	<b>66.7</b>	<b>51.5</b>	44.9	<b>55.7</b>	<b>99.6</b>	<b>97.1</b>	62.4	<b>93.8</b>	92.7	<b>97.4</b>	<b>89.8</b>	88.3	<b>93.9</b>
Chisquare	-0.9	16.1	10.3	16.4	15.9	15.2	15	15.4	7.6	19.6	45.5	26.1	43.7	39.7	42.8	38.6	40.7	25.1	49.4
	-0.5	18.6	10.9	16.6	16.8	17.1	16.4	16.5	8.4	19.6	52.8	29.6	47.1	43.6	47.9	44.9	43.5	29.3	52
	-0.1	23.2	13.1	17.2	17.3	20	19.7	17.4	10.4	21.1	64.1	37.4	50.2	49.5	54.7	53.9	47.8	35.5	56.4
	0.1	27	15.8	18	19.5	23.1	23.3	19.1	12.1	23.1	72.3	44	51.9	54.2	60.8	61.4	51.4	40.9	60.7
	0.5	40.3	22	16.8	24.2	30.1	32.3	22.1	16.6	27.5	90.4	65.5	57	70.3	77.7	79.8	62.1	58.4	74.2
	0.9	85.7	61.5	13.9	52.5	43.6	69	36.3	40.5	54.9	99.9	99.1	58.9	97.9	96.9	99.6	86.5	95.2	98.2

**Table 17** Power simulation results ( $\alpha = 0.05$ ) of the tests for different distributions under varying correlation values ( $\rho$ ) with sample sizes  $n = 30$  and heteroscedastic covariance matrix  $\Sigma_2$  under the MAR framework. Values of too liberal tests corresponding to red values in the  $\Sigma_2$ -column from Table 5 are printed in red colour.

Dist	$\rho$	$\delta = 0.5$								$\delta = 1$									
		$F$	Parametric bootstrap				Alternatives				$F$	Parametric bootstrap				Alternatives			
			$T_W^*$	$T_A^*$	$T_M^*$	$T_L^*$	$T_t$	$T_N$	$T_F$	$T_P$		$T_W^*$	$T_A^*$	$T_M^*$	$T_L^*$	$T_t$	$T_N$	$T_F$	$T_P$
Normal	-0.9	20.1	13.6	20.4	20.3	19.6	17.9	17.3	7.7	24.7	63.1	43.3	60.9	59.1	60.4	53.6	52.4	35.2	67
	-0.5	24.6	15.6	22.3	22.8	22.1	21	18.8	9.6	26.4	71.8	49.8	64.9	63.4	65.5	61.4	56.3	42	70.6
	-0.1	31.9	18.5	23.5	24.8	26.1	25.7	20.9	11.6	28.5	82.8	59.2	69.4	70.1	73.1	71.4	62.2	50.8	75.3
	0.1	37.3	21.1	24.7	27.2	29.7	29.6	22.3	14.3	31.1	89.5	66.9	72.6	75.4	79.9	79.7	66.3	58.5	80.5
	0.5	55.9	31.9	24.5	34.9	42.1	43.1	27	21.8	38.2	98.6	87.3	77.8	89.6	93.7	94.2	77.9	79.5	91.3
	0.9	97.6	83.6	23.2	77	70.8	89.8	49.8	64.1	78.7	100	100	84.9	100	100	100	98	100	100
Laplace	-0.9	21.4	14.8	21.4	24.3	20.6	18.4	22.4	10.3	26.6	62.4	46	60.8	61.2	60.1	52.7	64.1	42.3	68
	-0.5	26.4	16.6	24	25.6	23.6	21	26.3	12.7	29.1	72.4	53.7	66.8	67.9	66.8	61.1	71.6	52	73.2
	-0.1	32.6	19.2	24.9	27.4	26.8	25.5	29.8	16.3	31.3	82.9	62.5	72	73.7	74.2	71.7	78.1	62.7	78
	0.1	37.8	22.1	26.8	29.9	30.9	29.3	33	19.4	33.7	88.6	69.3	75.1	78.6	79.8	78.8	81.3	69.4	82.1
	0.5	57.9	33.6	28	38.5	43.8	44.5	40.7	30.6	43	98.4	86.6	80.3	89.7	93.2	92.9	89.6	85.1	91.6
	0.9	96.9	83.1	26	76.7	72.2	87.7	66	72	78.2	100	100	85.7	99.9	100	100	98.7	99.8	99.9
Exponential	-0.9	24.6	16.5	24.8	27.3	24	21.6	39.1	17.1	29.2	63.6	48.6	61.3	63.9	61.5	54.8	75.7	50.7	67.7
	-0.5	28.9	18	26.2	27.9	26.9	25.3	42.9	19.4	29.1	71.5	51.1	66.4	68	66.4	63	79.9	57.2	70.6
	-0.1	35.8	23.2	29.2	30.7	32.8	32.2	48.9	24.6	32.5	81	59.3	70.2	71.6	72.1	72	84.1	64	74.4
	0.1	41	26.5	30	32.1	35.6	35.9	50.6	27.5	34.2	86.8	66.1	73.5	75.6	77.4	78.2	87.3	69.5	78.3
	0.5	58.2	38.2	29.8	39.7	46.5	48.3	56.5	36	42	96	82.2	77	84	88.3	89.4	90.8	80.7	85.9
	0.9	92	80	23.6	69.3	68.7	79.8	71.2	64.5	69.2	100	99.8	82.1	98.9	99.3	99.7	98.2	98.2	98.9
Chisquare	-0.9	21.2	14	20.8	21.6	20	18.4	19.6	8.8	25.5	60.8	43.5	59.6	58.3	59.2	53.6	55.8	37.2	65.9
	-0.5	25.8	15.8	23.4	23.2	23.4	22.8	21.5	10.6	27.2	71.7	49.2	65.1	64.2	65.8	61.7	61.1	44.4	70
	-0.1	32.2	19.3	23.8	24.7	26.8	26.9	23.7	13.5	28.3	82.7	58.2	69.2	69.7	73	71.6	66.1	51.9	75.2
	0.1	38.1	22.2	24.9	27.2	31.5	32	25.9	15.7	31.1	88.8	66.1	72.4	74.4	78.6	78.9	69.7	59.6	79
	0.5	56.3	33.1	24.3	35.8	43	44.8	31.2	23.8	38.9	98.4	85.9	78.3	88.1	93	93.8	81.2	79	90.4
	0.9	96.4	82	21.3	73.7	69.3	86.5	53.3	62.4	75	100	100	83.9	99.9	99.9	100	97.4	99.7	99.9

**Table 18** Type-I error and power simulation results ( $\alpha = 0.05$ ) of Little's test  $T_L$  for different distributions under varying correlation values ( $\rho$ ) with different sample sizes and homoscedastic covariance matrix  $\Sigma_1$  under the MCAR framework.

Dist	$\rho$	$(n_c, n_u) = (10, 10)$			$(n_c, n_u) = (10, 30)$			$(n_c, n_u) = (30, 10)$		
		$\delta = 0$	$\delta = 0.5$	$\delta = 1$	$\delta = 0$	$\delta = 0.5$	$\delta = 1$	$\delta = 0$	$\delta = 0.5$	$\delta = 1$
Normal	-0.9	4.7	16.6	53.2	3.8	24.4	77.8	5.1	35.9	88.1
	-0.5	5.2	19.2	55.9	5.1	24.5	71.4	4.8	39.8	93.2
	-0.1	5.2	21.4	63	5.8	25	71.4	4.9	49.4	97.2
	0.1	5.9	23.9	68.5	6.3	28.5	75.1	5.6	56.6	98.7
	0.5	6.5	35.7	86.1	7.1	38.4	88.3	5.2	78.7	100
	0.9	12	85.7	99.9	8.5	89	100	11.6	99.9	100
Laplace	-0.9	4.2	17.8	55	3.3	27	78.3	5.4	35.6	87.2
	-0.5	4.8	20.3	58.6	3.9	24.8	70.8	5.3	41.1	91.7
	-0.1	5.1	23.7	65.5	5	28.1	73	5	50.2	96.6
	0.1	5.3	26.8	70.2	5.4	30.6	76.7	5.8	58.2	98.4
	0.5	6.1	40	86.5	6.1	42.8	88.4	5.4	79.2	99.9
	0.9	12.4	83.4	99.2	9.4	88.4	100	14.3	99.3	100
Exponential	-0.9	4.2	20.2	57.3	3.8	32.1	81.2	5.4	37.8	87.6
	-0.5	5.2	24.6	61.7	5.6	35.3	74.9	4.8	42.8	91.7
	-0.1	6.3	29.8	66.4	7.7	36	71.6	4.8	52.3	95.6
	0.1	7	33.2	69.5	8.7	38.3	74	5.6	58.6	97.2
	0.5	9.8	43.7	83.2	10.6	44.8	81.8	6.2	78.2	99.9
	0.9	15.3	82.9	98.2	15.9	85.8	99.7	15.7	98.7	100
Chisquare	-0.9	4.6	18.1	54.5	3.3	27.4	79.1	5.4	36	87.7
	-0.5	4.8	19.9	56.6	4.8	26.4	70.8	5.3	41.3	93
	-0.1	5.6	23	63.8	5.9	27.4	71.1	5.1	50.7	97.2
	0.1	5.8	25.5	67.6	6	29.7	73.7	5.1	57.1	98.7
	0.5	6.8	36.3	85.6	6.5	39.1	86.5	5.1	78.5	100
	0.9	12.3	84.9	99.8	8.8	88.1	100	12.6	99.8	100

**Table 19** Type-I error and power simulation results ( $\alpha = 0.05$ ) of Little's test  $T_L$  for different distributions under varying correlation values ( $\rho$ ) with different sample sizes and heteroscedastic covariance matrix  $\Sigma_2$  under the MCAR framework.

Dist	$\rho$	$(n_c, n_u) = (10, 10)$			$(n_c, n_u) = (10, 30)$			$(n_c, n_u) = (30, 10)$		
		$\delta = 0$	$\delta = 0.5$	$\delta = 1$	$\delta = 0$	$\delta = 0.5$	$\delta = 1$	$\delta = 0$	$\delta = 0.5$	$\delta = 1$
Normal	-0.9	4.8	12.9	38.4	3.9	17.6	59.1	5.2	26.1	73.8
	-0.5	5.5	14.1	40.2	5.5	16.8	49.7	4.9	28.5	79.7
	-0.1	5.7	15.9	44	6.3	17.4	48	5	34.4	87.3
	0.1	6	16.9	48	6.6	19	52.1	5.3	40.1	92
	0.5	6.8	23.9	64.8	7.5	24.4	65.6	5.3	57.5	98.8
	0.9	12.4	62.3	97.4	8.1	63.3	99.2	11.1	95.8	100
Laplace	-0.9	4.4	13.5	41.8	3.6	19.5	60.5	5.3	26.4	73.1
	-0.5	5.2	15	42.8	4.1	17.5	51.6	5.4	29.6	78.5
	-0.1	5.3	17.3	48.4	5.5	19.6	52.8	5.2	35.6	86.9
	0.1	5.5	19	52.6	5.8	20.5	56.5	6	41.8	91.1
	0.5	6.3	27.8	69.1	6.3	28.3	69.8	5.2	59.4	98.5
	0.9	13.6	64.2	94.8	8.6	66.9	98.1	14.5	93.2	100
Exponential	-0.9	5	18.6	46	4.9	26.5	65.6	5.7	30.4	73.1
	-0.5	6.4	21.6	49.4	7.4	28.8	60.3	5.2	32.9	78.5
	-0.1	7.6	25.4	52.5	9.4	29.2	56.3	5.1	39.4	84.4
	0.1	8.9	27.8	54.6	10.8	30.8	58.1	6	44.3	87.5
	0.5	11.5	34.6	66.1	11.9	33.6	64.7	7.2	59.5	96.1
	0.9	18.1	65.2	91.1	15.2	65.1	95.4	17.2	88.3	99.6
Chisquare	-0.9	4.8	14.7	40.8	3.6	20.3	61.3	5.5	27.1	73.1
	-0.5	5.1	15.8	41.5	5.3	18.9	51.6	5.3	30.5	79.7
	-0.1	6.1	17.4	45.6	6.4	19.8	49.9	5.4	36.2	87
	0.1	6.2	19	49	6.6	21	52.1	5.2	41.2	91
	0.5	6.9	25.6	64.6	7	26.2	64.3	5.3	57.4	98.3
	0.9	12.8	61.7	95.8	8.4	62.9	98.5	11.4	94.2	100

**Table 20** Type-I error and power simulation results ( $\alpha = 0.05$ ) of Little's test  $T_L$  for different distributions under varying correlation values ( $\rho$ ) with different sample sizes and homoscedastic covariance matrix  $\Sigma_1$  under the MAR framework.

Dist	$\rho$	$n = 10$			$n = 20$			$n = 30$		
		$\delta = 0$	$\delta = 0.5$	$\delta = 1$	$\delta = 0$	$\delta = 0.5$	$\delta = 1$	$\delta = 0$	$\delta = 0.5$	$\delta = 1$
Normal	-0.9	4.9	10.1	25.3	5.9	19.4	57.3	5.1	28	76.6
	-0.5	5.5	11.8	27.8	5.2	22.1	65.3	5	31.3	82.9
	-0.1	6.3	14.6	36.3	5.2	27.2	75.3	5.2	39	90.4
	0.1	6.7	16.4	43	5.5	30.9	81.6	5.5	43.5	94.6
	0.5	8.1	25.3	65.6	6.4	46.1	96	5.8	63.8	99.6
	0.9	12.8	72.8	98.1	14	93.6	100	12.8	98.9	100
Laplace	-0.9	5	12.2	30.1	4.9	21.3	58.8	5.1	29.3	76.8
	-0.5	5.4	13.5	34.4	5.2	23.6	66	5	32.6	83.2
	-0.1	5.7	15.7	43.5	5	28.4	76.2	5	39.5	90.2
	0.1	6.8	19.2	49.5	5.2	31.8	81.6	5.3	45.3	93.4
	0.5	8.7	31.3	71.3	6.2	50.6	94.9	6	64.6	99.2
	0.9	12.4	75.4	96.3	14.3	89.5	99.7	14.1	96.8	100
Exponential	-0.9	5.1	12.6	34.4	5.4	22.6	61.7	5.1	31	76.4
	-0.5	5.3	14.7	38.7	5.2	26.3	68.5	5.6	34.9	82.3
	-0.1	6.5	19	46.9	5.2	32.4	76.4	5.1	42.5	88.7
	0.1	7.1	22.9	54.2	5.4	37.7	80.7	5.4	47.9	92.4
	0.5	10.6	35.2	72.5	7.6	50.9	93.3	6.8	65.3	98.7
	0.9	11.2	74.4	96	13.7	89.6	99.1	14.6	95.9	99.9
Chisquare	-0.9	5.5	10.3	25.9	5.4	19.7	58.1	4.9	27.2	76.4
	-0.5	5.3	12.2	30.6	5.4	22.2	66.1	5.2	31.5	82.6
	-0.1	6.3	14.8	37.9	5.2	27.8	74.4	5.4	38	90.9
	0.1	6.9	17	44.8	5.3	31.7	80.9	5.1	44.1	94
	0.5	8	26	66.3	6.2	47.1	95.3	5.4	65	99.4
	0.9	13.4	74.3	97.9	14.3	93.7	100	13.3	98.9	100

**Table 21** Type-I error and power simulation results ( $\alpha = 0.05$ ) of Little's test  $T_L$  for different distributions under varying correlation values ( $\rho$ ) with different sample sizes and heteroscedastic covariance matrix  $\Sigma_2$  under the MAR framework.

Dist	$\rho$	$n = 10$			$n = 20$			$n = 30$		
		$\delta = 0$	$\delta = 0.5$	$\delta = 1$	$\delta = 0$	$\delta = 0.5$	$\delta = 1$	$\delta = 0$	$\delta = 0.5$	$\delta = 1$
Normal	-0.9	5.9	8.6	18.8	5.3	15	43.2	5.6	20.2	61.2
	-0.5	5.8	9.4	21.2	5.3	17.1	47.8	5.2	22.5	65.9
	-0.1	6.6	11.5	26.7	5.6	19.4	56.5	4.9	26.9	74.1
	0.1	6.6	13.3	30.5	6.1	22.6	63.1	5.7	31.2	80.9
	0.5	8.6	18.1	46.4	5.7	31.4	82.1	5.4	43.8	94.3
	0.9	16.3	55.3	91.2	14.3	74.8	99.1	11.7	88.8	100
Laplace	-0.9	5.6	10.3	22.7	5.6	15.7	45.9	5.6	21.4	61.1
	-0.5	5.5	11.1	25.9	5.2	17.6	51.3	5.1	24.3	67.4
	-0.1	6	12.5	31.6	5.2	21.2	59	5	28	74.9
	0.1	6.8	14.7	36.9	6	23.4	65	5.1	32	80.6
	0.5	8.9	23.4	54.7	6.6	35.1	82.2	5.3	46.2	93.6
	0.9	14.9	60.9	90.8	16.3	73.8	97.3	14.9	85.4	99.7
Exponential	-0.9	5.9	14.2	30	5.8	20.7	48.6	5.8	25	62.5
	-0.5	6	15.6	32.2	5.5	23.5	53.3	6	27	67
	-0.1	7.3	18.6	39	6.5	27.4	61.2	6	33.3	73
	0.1	8.4	22	43.8	6.7	30	65.1	6.3	36.6	78.5
	0.5	13.2	31	59.2	8.6	40.1	79.4	7	48.7	90
	0.9	17.1	64	90.4	18.5	71.8	93.8	17	80.5	98.2
Chisquare	-0.9	6.5	9.4	20.7	5.2	16.5	44.6	5.4	21	60
	-0.5	5.8	10.8	22.7	5.4	17.8	49.2	5.3	23.7	66.2
	-0.1	6.8	12.7	28.5	5.9	21.2	56.8	5.6	27.7	74
	0.1	7.2	14.7	32.6	5.9	24.6	63.3	5.6	32.3	79.9
	0.5	8.8	21.6	50.1	6.2	33.6	80.8	5.7	45.2	93.7
	0.9	16.8	58.8	91.1	15.4	73.3	98.3	12.6	86.5	99.9

**Table 22** Type-I error simulation results ( $\alpha = 0.05$ ) of the tests for different distributions under varying correlation values ( $\rho$ ) with sample sizes  $(n_c, n_u) = (16, 74)$  and homoscedastic and heteroscedastic covariance matrices  $\Sigma_1$  and  $\Sigma_2$  respectively under the MCAR framework.

Dist	$\rho$	$\Sigma_1$								$\Sigma_2$									
		F	Parametric bootstrap				Alternatives				F	Parametric bootstrap				Alternatives			
			$T_W^*$	$T_A^*$	$T_M^*$	$T_L^*$	$T_t$	$T_N$	$T_F$	$T_P$		$T_W^*$	$T_A^*$	$T_M^*$	$T_L^*$	$T_t$	$T_N$	$T_F$	$T_P$
Normal	-0.9	5.1	5.5	5.6	6.7	5.5	5.3	6.6	5	9.1	5.3	5.6	5.6	7	5.7	5.5	7.2	6.5	9.6
	-0.5	5.1	5.1	5.1	6	5.1	4.9	5.9	4.2	8.2	5.1	4.9	5.1	6.2	5.1	4.9	6.3	5.9	8.8
	-0.1	5	5.2	5.6	6.2	5.4	4.8	6	4.4	7.5	5.1	5.2	5.8	6.4	5.6	5	7	6.4	8.5
	0.1	5	5.2	5.3	5.6	5.2	4.9	5.8	4.1	7	4.9	5.1	5.3	5.7	5.2	5.1	6.8	5.8	8.1
	0.5	5.2	5.5	5.4	5.7	5.5	5.4	6.1	4.3	6.3	5.2	5.5	5.5	6.2	5.5	5.3	7	6.2	7.8
	0.9	5	5.3	5.3	4.7	4.8	5.2	5.8	4.3	5.3	5.1	5.3	5.7	7.1	4.9	4.9	6.8	5.7	7.9
Laplace	-0.9	5.3	5	5.2	6.2	4.9	5	6.4	4.9	8.6	5.3	5	5.1	6.2	5.1	5	6.6	6.2	9.2
	-0.5	5.5	5	5.3	6.3	5.3	5	6.2	4.6	8.6	5.4	4.8	5.2	6.3	5.1	5	6.6	5.9	9.3
	-0.1	5.2	4.2	4.8	5.5	4.8	4.7	5.9	4.2	7	5.3	4.2	4.8	5.9	4.7	4.4	6.6	5.4	8.3
	0.1	5.6	4.8	5.3	6	5.5	5.2	6	4.3	7.3	5.3	4.8	5.3	6.4	5.3	5.1	6.6	5.6	8.3
	0.5	5.1	5	5	5.8	5.2	5.1	5.9	4.4	6.6	5.1	4.8	5.2	6.5	5.2	4.9	6.6	5.5	8
	0.9	5.2	4.5	5.3	4.4	4.3	4.7	5.4	4.4	5.3	5.1	4.5	5.3	7.4	4.4	4.7	6.1	5.5	7.6
Exponential	-0.9	5	5.5	4.8	6.2	5.9	4.7	6.1	4.6	8.6	5.1	6.2	5.3	7.1	6.6	5	10.3	4.9	9.8
	-0.5	5.3	6.5	5.2	7.2	7	4.5	5.7	4.1	7.8	5.5	7.2	6	7.8	8	4.9	10.7	5.4	9.2
	-0.1	5.2	7.7	5.8	8.2	7.9	4.8	6.1	4.3	7.6	5.3	8.3	7	9.1	8.7	5.4	12.4	6.4	9.5
	0.1	5	8.2	5.9	8.4	8	4.7	5.9	4.1	7.1	5.2	8.8	7.5	9.4	8.7	5.2	12.3	6.3	9.1
	0.5	5.4	8.8	6.5	7.9	9.4	4.5	5.4	3.8	6.1	5.7	9.2	8.2	9.5	9.6	5.5	12.8	5.8	8.8
	0.9	5.1	9.2	6.5	5	9.1	4.4	5.5	4.4	5	6	9.2	7.8	11.5	8.7	7.7	16.1	6.2	10.1
Chisquare	-0.9	5.5	5.4	5.2	6.3	5.4	5.1	5.9	4.6	8.6	5.3	5.5	5.1	6.5	5.4	4.9	6.8	5.4	9
	-0.5	5.4	5.7	5.5	6.4	5.7	5.1	6.1	4.4	8.4	5.3	5.6	5.6	6.4	5.8	5.1	7	5.7	9
	-0.1	5.2	5.3	5.6	5.9	5.5	4.9	6.4	4.5	7.4	5.1	5.2	5.5	6	5.5	5	7.3	5.5	8.6
	0.1	5.6	5.4	5.2	5.8	5.5	4.7	5.6	4	6.8	5.5	5.4	5.2	6	5.4	4.8	6.8	5.2	8
	0.5	5.2	5	5.4	5.6	5.7	4.9	6	3.9	6.3	5.4	5.2	5.7	6.2	5.6	5.1	7.3	5	7.8
	0.9	5.2	5.7	5.5	4.4	4.9	4.7	5.4	4.5	5.1	5.3	5.6	5.6	7.5	5.1	5.1	7.7	5.3	8



**Table 23** Type-I error simulation results ( $\alpha = 0.05$ ) of the tests for different distributions under varying correlation values ( $\rho$ ) inspired from the dependency structure of the genes data example with sample sizes  $(n_c, n_u) = (16, 74)$  and homoscedastic and heteroscedastic covariance matrices  $\Sigma_1$  and  $\Sigma_2$  respectively under the MCAR framework.

Dist	$\rho$	$\Sigma_1$								$\Sigma_2$									
		F	Parametric bootstrap				Alternatives				F	Parametric bootstrap				Alternatives			
			$T_W^*$	$T_A^*$	$T_M^*$	$T_L^*$	$T_t$	$T_N$	$T_F$	$T_P$		$T_W^*$	$T_A^*$	$T_M^*$	$T_L^*$	$T_t$	$T_N$	$T_F$	$T_P$
Normal	-0.3	5.1	5.8	5.9	6.4	5.7	5.3	6.6	4.8	8.2	5.3	5.7	5.8	6.6	5.7	5.4	7.2	6.5	8.9
	-0.1	5.1	5.1	5.3	5.7	5	4.9	5.7	4	7.3	5	5	5.2	5.9	5	4.8	6.3	5.7	8.5
	0.3	5	5.2	5.6	5.8	5.3	4.8	6.1	4.2	6.7	5	5.1	5.8	6.1	5.6	5.1	6.8	6	8
	0.4	5	5.2	5.2	5.4	5.2	4.9	5.7	4.1	6.4	4.9	5.2	5.1	5.8	5.3	5.1	6.6	5.8	7.7
	0.7	5.2	5.5	5.6	5.6	5.6	5.4	6	4.5	6	5.2	5.5	5.8	6.5	5.4	5.3	7	6.1	7.8
	0.8	5	5.2	5.4	5.1	5.3	5.2	5.8	4	5.6	5	5.2	5.6	6.2	5.2	5.1	6.7	5.7	7.5
Laplace	-0.3	5.3	4.8	5.3	5.9	5.2	5	6.1	4.4	7.6	5.2	4.7	5	6	5.1	5	6.6	5.7	8.7
	-0.1	5.5	4.9	5.4	6.2	5.1	5	6.3	4.4	8	5.3	4.9	5.4	6.5	5.1	4.9	6.5	5.8	8.9
	0.3	5.2	4.2	4.7	5.5	4.7	4.7	5.9	4.1	6.3	5.3	4.2	4.7	6.1	4.6	4.4	6.6	5.4	7.7
	0.4	5.6	4.8	5.1	5.7	5.2	5.2	6	4.3	6.8	5.3	4.6	5.2	6.4	5.1	5	6.6	5.8	8.1
	0.7	5.1	5	4.9	5.5	5.3	5.1	5.9	4.4	6.2	5.2	4.9	5	6.5	5	4.9	6.7	5.6	7.8
	0.8	5.2	4.6	4.9	5	4.9	4.7	5.5	4.4	5.6	5.1	4.4	5.2	6.7	4.8	4.7	6.1	5.6	7.6
Exponential	-0.3	5	7.5	5.3	7.7	7.6	4.7	6	4.4	7.7	5	8.1	6.5	8.7	8.5	5	11.3	5.9	9.6
	-0.1	5.3	7.7	5.6	7.7	7.3	4.5	5.6	3.9	7.1	5.6	8.2	6.7	8.5	8	5	12.1	6	9
	0.3	5.2	8.5	6.4	8.4	9	4.8	6	4.2	7	5.4	9	8	9.6	9.5	5.7	13.1	6.6	9.2
	0.4	5	8.7	6.3	8.3	8.9	4.7	5.9	4	6.4	5.3	9.1	8.1	9.2	9.6	5.7	12.4	6.2	8.9
	0.7	5.4	9.1	6.6	7.2	9.9	4.5	5.3	4.1	5.6	5.8	9.3	8.2	9.8	10	6.3	13.6	5.9	9
	0.8	5.2	8.9	6.3	6.5	9.7	4.4	5.4	4.2	5.2	5.8	9.3	8.2	10.3	9.8	6.7	14.3	6.2	9
Chisquare	-0.3	5.4	5.2	5.4	5.9	5.5	5.1	5.9	4.3	7.5	5.3	5.4	5.3	6.1	5.7	5	6.8	5.2	8.4
	-0.1	5.4	5.7	5.7	6.1	5.7	5.1	6.2	4.1	7.7	5.3	5.8	5.8	6.2	5.7	5.1	7.1	5.4	8.4
	0.3	5.2	5.2	5.3	5.7	5.6	4.9	6.2	4.1	6.5	5.1	5.1	5.5	6	5.5	5	7.6	5.2	7.9
	0.4	5.6	5.4	5.1	5.5	5.3	4.7	5.5	4	6.2	5.5	5.4	5.3	5.9	5.3	4.8	6.9	5.2	7.6
	0.7	5.3	5.1	5.3	5.4	5.7	4.9	5.9	4	5.7	5.3	5.2	5.6	6.4	5.6	5.2	7.4	5.1	7.8
	0.8	5.2	5.5	5.2	5.1	5.5	4.7	5.4	4.1	5.4	5.3	5.6	5.6	6.7	5.5	5	7.3	5.3	7.7

**Table 24** Adjusted two-sided P-values of the breast cancer study based on Holm's method.

Gene	Parametric bootstrap				Alternatives			
	$T_{\mathbf{W}}^*$	$T_{\mathbf{A}}^*$	$T_{\mathbf{M}}^*$	$T_{\mathbf{L}}^*$	$T_t$	$T_{\mathbf{N}}$	$T_{\mathbf{F}}$	$T_{\mathbf{P}}$
TP53	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
ABCC1	0.016	0.021	0.016	0.016	1.000	0.021	0.032	0.000
HRAS	0.049	0.016	0.021	0.016	0.176	0.008	0.032	0.000
GSTM1	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
ERBB2	0.258	0.144	0.066	0.084	0.816	0.426	0.414	0.042
CD8A	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
C1D	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
GBP3	0.98	1.000	1.000	0.42	0.581	0.515	1.000	1.000

## References

- Brunner, E. and Puri, M. L. (2001). Nonparametric methods in factorial designs. *Statistical papers* **42**, 1–52.
- Friedrich, S. and Pauly, M. (2018). Mats: Inference for potentially singular and heteroscedastic manova. *Journal of Multivariate Analysis* **165**, 166–179.
- Graybill, F. A. (1976). *Theory and application of the linear model*, volume 183. Duxbury press North Scituate, MA.
- Konietschke, F., Bathke, A. C., Harrar, S. W. and Pauly, M. (2015). Parametric and nonparametric bootstrap methods for general manova. *Journal of Multivariate Analysis* **140**, 291–301.
- Little, R. J. A. and Rubin, D. B. (2014). Incomplete data. *Wiley StatsRef: Statistics Reference Online* .
- Rao, C. R., Mitra, S. K. et al. (1972). Generalized inverse of a matrix and its applications. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Theory of Statistics*. The Regents of the University of California.
- Rubin, D. B. (2004). *Multiple imputation for nonresponse in surveys*, volume 81. John Wiley & Sons.