UNIVERSITY OF BELGRADE
FACULTY OF MATHEMATICS

Danijel G. Aleksić

U- AND *V*-STATISTICS FOR INCOMPLETE DATA AND THEIR APPLICATION TO MODEL SPECIFICATION TESTING

Doctoral Dissertation

Belgrade, 2026

УНИВЕРЗИТЕТ У БЕОГРАДУ
МАТЕМАТИЧКИ ФАКУЛТЕТ

Данијел Г. Алексић

U- И *V*-СТАТИСТИКЕ ЗА НЕКОМПЛЕТНЕ ПОДАТКЕ И ЊИХОВА ПРИМЕНА У ТЕСТИРАЊУ САГЛАСНОСТИ СА МОДЕЛОМ

Докторска дисертација

Београд, 2026.

Advisor:

Professor Bojana Milošević, PhD
Associate Professor at the University of Belgrade, Faculty of Mathematics

Dissertation defense committee members:

Marija Cuparić, PhD
Assistant Professor at the University of Belgrade, Faculty of Mathematics

Marko Obradović, PhD
Associate Professor at the University of Belgrade, Faculty of Mathematics

Apostolos Batsidis, PhD
Associate Professor at the University of Ioannina (Greece), Department of Mathematics

Acknowledgments

First of all, I am grateful to my late mother Milofinka, who had a key influence on my early education and who taught me to read and write at the age of three. She was happier about every success of mine than I ever was, and I am sure that this success is no exception, only this time I cannot personally witness her joy. I am grateful to my father Gojko, who has always been my support and protection in life, and who even today is able to fulfill all my wishes. He often says: *"My father Dragoje was a road worker, and I am a truck driver. Who could have imagined that Dragoje's grandson would become a Doctor of Science!"* I am glad that the effort of my ancestors was not in vain.

I would like to thank my advisor, Prof. Dr. Bojana Milošević, who is a true example of how a supervisor should guide their students. I owe my development as a statistician to her and to her fine sense of when to give criticism and when to give praise.

I am grateful to the members of the committee, Dr. Marija Cuparić, Prof. Dr. Marko Obradović, and Prof. Dr. Apostolos Batsidis, for carefully reading this dissertation and for the many useful suggestions that made it much better than it was at the beginning.

My thanks also go to my colleagues from the Department of Mathematics at the Faculty of Organizational Sciences, who understood my scientific duties, which often meant that other work had to wait.

I am thankful to my friends who believed in me when I did not. My thanks go to Mr. Rami H. Haider, who always knew how to listen and who perfectly pretended to be interested in missing data analysis. I am also grateful to Mr. Dalibor N. Danilović for giving me an example of what it means to have patience. I thank Dr. Žikica G. Lukić for our countless debates, which taught me that conviction is no substitute for clarity.

My gratitude also goes to all the teachers who taught me during my schooling. I am especially thankful to those who taught me mathematics: Milan Rankić, my elementary school teacher from the first to the fifth grade; then Gospava Vidović and Novica Đerić, who taught me from the sixth to the ninth grade. I am most grateful to my high school teacher, Ivan Pavlović. He may not be the only one responsible for the fact that I am a mathematician today, but he is surely the reason why I am a teacher.

I am also thankful to everyone who has helped me in ways I may not have even noticed.

Above all, I thank God for the talent He gave me. I hope I am using it in the right way.

Dissertation title: U- and V-statistics for incomplete data and their application to model specification testing

Abstract: This dissertation addresses the problem of model specification testing in situations where data are incomplete, utilizing the existing theory of non-degenerate and weakly degenerate U- and V-statistics. The first two chapters lay the theoretical groundwork by presenting essential concepts related to U- and V-statistics and the general mathematical framework of missing data analysis, which serve as the foundation for the new results developed in subsequent chapters.

In Chapter 3, a novel test for assessing the missing completely at random (MCAR) assumption is introduced. This test demonstrates improved control of the type I error rate and superior power performance compared to the main competitor across the majority of the simulated scenarios examined.

Chapter 4 explores the application of Kendall's test for independence in the presence of MCAR data. It provides both theoretical insights and simulation-based comparisons of the complete-case analysis and median imputation, pointing out their individual advantages and drawbacks.

Chapter 5 focuses on testing for multivariate normality when data are incomplete. It rigorously establishes the validity of the complete-case approach under MCAR and proposes a bootstrap method to approximate p-values when imputation is employed. Additionally, various imputation techniques are evaluated with respect to their impact on the type I error and the power of the test.

Finally, Chapter 6 adapts the energy-based two-sample test to handle missing data by introducing a weighted framework that makes full use of all available observations. Alongside some theoretical developments, the chapter presents two distinct bootstrap algorithms for p-value estimation under this approach. Additionally, the performance of several imputation methods is examined in this context, and appropriate bootstrap algorithm is proposed for that setting.

Keywords: missing data, model specification testing, tests of MCAR, independence testing, goodness-of-fit testing, two-sample testing, bootstrap.

Scientific Area: Mathematics

Scientific Sub-Area: Probability and Statistics

Наслов докторске дисертације: U- и V-статистике за некомплетне податке и њихова примена у тестирању сагласности са моделом

Сажетак: Ова дисертација бави се проблемом тестирања сагласности са моделом у присуству недостајућих података, ослањајући се на постојећу теорију недегенерисаних и слабо дегенерисаних U- и V-статистика. Прве две главе постављају теоријске основе тако што приказују основне концепте у вези са U- и V-статистикама, као и општи математички оквир за анализу недостајућих података, што служи као полазна тачка за нове резултате из наредних глава.

Глава 3 уводи нови тест за тестирање претпоставке да подаци недостају на потпуно случајан начин (MCAR). Испоставља се да нови тест у већини проучаваних сценарија има бољу контролу грешке прве врсте у односу на главног конкурента, а такође има боље перформансе и у смислу моћи теста.

У Глави 4 анализира се примена Кендаловог теста независности на податке који недостају на потпуно случајан начин. Приступи уклањања свих некомплетних елемената узорка и попуњавања узорачком медијаном пореде се како теоријски, тако и емпиријски, указујући на предности и недостатке сваког приступа.

Циљ Главе 5 јесте тестирање претпоставке вишедимензионалне нормалности података онда када они делимично недостају. Теоријски се потврђује валидност приступа уклањања некомплетних елемената узорка при претпоставци MCAR недостајања, а бутстреп алгоритам се предлаже за апроксимацију p-вредности када се подаци попуњавају неким од метода. Такође, разни методи попуњавања пореде се у смислу утицаја на грешку прве врсте и моћ теста.

Коначно, Глава 6 бави се тестом једнакости у расподели заснованом на *energy* растојању и његовом прилагођавању условима недостајућих података. Уводи се тежински приступ који је у стању да искористи све доступне податке. Поред теоријских резултата, нуде се два различита бутстреп алгоритма за апроксимацију p-вредности при том приступу. Такође, проучавају се и перформансе неколицине метода попуњавања недостајућих података у контексту овог теста и предлаже се одговарајући бутстреп алгоритам.

Кључне речи: недостајући подаци, тестирање сагласности са моделом, тестирање потпуно случајног недостајања, тестирање независности, тестирање сагласности са расподелом, тестирање једнакости у расподели, бутстреп.

Научна област: Математика

Ужа научна област: Вероватноћа и статистика

Contents

Table 1: Abbreviations used throughout the dissertation and their meanings

Abbreviation	Meaning
BHEP	Baringhaus–Henze–Epps–Pulley
CDF	Cumulative distribution function
ECDF	Empirical CDF
EMAR	Everywhere MAR
EMCAR	Everywhere MCAR
GOF	Goodness-of-fit
IID	Independent and identically distributed
MAAR	Missing always at random
MAR	Missing at random
MACAR	Missing always completely at random
MCAR	Missing completely at random
MLE	Maximum likelihood estimate
MNAAR	Missing not always at random
MNAR	Missing not at random
MVN	Multivariate normality
RMAR	Realized missing at random
RMCAR	Realized missing completely at random

Table 2: Mathematical notation used throughout the dissertation

Symbol	Meaning
$\mathbb{N} = \{1, 2, \dots\}$	The set of natural numbers
$\mathbb{R}$	The set of real numbers
I_d	Identity $d \times d$ matrix (subscript omitted when clear)
J_d	$d \times d$ matrix with all elements equal to 1
$\mathbb{E}$	Expected value
$\mathbb{V}\mathrm{ar}$	Variance
$\mathbb{C}\mathrm{ov}$	Covariance (or covariance matrix)
$\mathbb{C}\mathrm{or}$	Pearson's correlation (coefficient or matrix)
$\mathbb{P}$	Probability
$\xrightarrow{D}$	Convergence in distribution
$\xrightarrow{P}$	Convergence in probability
A^T	The transpose of the matrix A
$\mathbb{T}\mathrm{r}$	The trace of a matrix/operator
$a \equiv b$	a is identically equal to b
:=, or =:	Equality by definition
$f = O_{\mathbb{P}}(g)$	$f = h \cdot g$ and h is bounded in probability
$f = o_{\mathbb{P}}(g)$	$f = h \cdot g$ and h has zero limit in probability
$\mathrm{sgn} x$	-1 if $x < 0$, 0 if $x = 0$, 1 if $x > 0$
$\mathcal{N}(\mu, \sigma^2)$	Univariate normal distribution with mean μ and variance σ^2
$\mathcal{E}(\lambda)$	Exponential distribution with rate parameter λ
$\mathcal{N}_d(\mu, \Sigma)$	d-variate normal distribution with mean μ and variance mat. Σ
Φ	CDF of the standard normal distribution

Preface

You are the handicap you must face.
You are the one who must choose your place.
You must say where you want to go,
How much you will study the truth to know.
God has equipped you for life, but He
Lets you decide what you want to be.

Edgar A. Guest, *Equipment*

One of the key assumptions in most statistical analyses is that the chosen model adequately reflects the data-generating process. Whether one is building a regression model, estimating parameters of a probability distribution, or examining the relationship between variables, an implicit assumption is always made: the model must be correctly specified. The existence of model specification testing plays an important role for the validity of any analysis. It asks, in a mathematically rigorous way, whether the assumptions we make about the data are justified enough by the data themselves.

Specification testing is a broad area, as it encompasses a wide range of fundamental problems in statistics. When assessing whether a particular distribution adequately describes the observed data, we enter the domain of goodness-of-fit testing. Determining whether two variables are independent, or whether a hidden dependence structure exists, falls under independence testing. Evaluating whether two samples are drawn from the same probability distribution leads us to two-sample testing. Additionally, problems such as testing whether data are missing according to a specific missingness mechanism are, at their core, tests of whether an assumed model is consistent with the observed data.

In many applied fields, model misspecification is not just an academic inconvenience, but can lead to serious problems. A notable example is the analysis of medical data. A model that underestimates risk, overlooks dependence, or assumes a normal distribution when none exists may yield results that are dangerously misleading. It is therefore no surprise that the development of specification tests has become a growing area of both theoretical and applied statistics.

However, real data are often subject to various forms of imperfection. Missing values, whether arising from nonresponse, measurement issues, or data corruption, are a challenge that not only fails to disappear over time, but becomes increasingly prominent nowadays as more and more data are becoming available. Classical specification tests, although elegant in theory, are often not applicable in the presence of missing data. This gap between the theoretical aspects of model specification testing and the practical obstacles of incomplete data has led to many new research efforts.

This thesis contributes to that work by developing a new testing procedure for assessing,

under a specified framework, whether data are missing completely at random (MCAR), a common assumption in many statistical analyses involving missing data. In addition to that, it proposes the adaptations of several popular specification tests, such as those for independence, goodness-of-fit, and equality of distributions, to remain applicable in the presence of missing data.

This work lies at the intersection of theory and practice, guided by the idea that formal statistical tools should still be useful even when the data are not perfect. Real-world data are often subject to missingness and other data imperfections, and if statistical methods are to be trusted, they need to work well in such settings. Although there is still much to explore and understand, the results in this thesis aim to take a step in that direction. They reflect an effort to make model specification testing more reliable and practical, and to connect classical statistical ideas with the real challenges encountered in modern data analysis.

Chapter 1

U- and *V*-statistics

This chapter introduces two important and fundamental classes of statistics, known as *U*- and *V*-statistics, which are natural generalizations of sample averages. *V*-statistics were first introduced by von Mises (1947), although not under that name, and *U*-statistics by Hoeffding (1948).

The literature on *U*- and *V*-statistics is extensive, and a comprehensive overview is beyond the scope of this thesis. In the remainder of this chapter, we present the definitions of *U*- and *V*-statistics, along with their asymptotic properties which are directly related or closely connected to the results discussed in Chapters 3–6 of this thesis. For further details, we refer the reader to Lehmann (1999) and Koroljuk and Borovskich (2010), as well as the references cited therein.

As we will see shortly, *U*- and *V*-statistics are defined very similarly to each other, which leads to asymptotic results that are often similar and, in some cases, identical.

1.1 Definitions

Let $X_1, \ldots, X_n$ be a sample of IID random vectors that take values in $\mathbb{R}^d$, and let $\phi(x_1, \ldots, x_m)$ be a measurable function symmetric with respect to its arguments. Then, a *U*-statistic with kernel ϕ is defined as

$$U_n = \frac{1}{\binom{n}{m}} \sum_{1 \leqslant i_1 < i_2 < \cdots < i_m \leqslant n} \phi(X_{i_1}, \ldots, X_{i_m}). \tag{1.1}$$

If we assume that $\mathbb{E}\big(\phi(X_1, \ldots, X_m)^2\big) < \infty$, we define σ_1^2 as

$$\sigma_1^2 = \mathbb{C}\mathrm{ov}\big(\phi(X_1, X_2, X_3, \ldots, X_m), \phi(X_1, X_2', X_3', \ldots, X_m')\big),$$

and

$$\sigma_2^2 = \mathbb{C}\mathrm{ov}\big(\phi(X_1, X_2, X_3, \ldots, X_m), \phi(X_1, X_2, X_3', \ldots, X_m')\big),$$

where X_j' is an independent copy of X_j. We say that U_n (or ϕ) is *non-degenerate* if $\sigma_1^2 > 0$, and *weakly degenerate* if $\sigma_2^2 > 0$ and $\sigma_1^2 = 0$. If we define the *first projection* of a kernel ϕ as

$$\phi_1(x) = \mathbb{E}\big(\phi(X_1, X_2, \ldots, X_m) \mid X_1 = x\big) - \theta,$$

where $\theta = \mathbb{E}\big(\phi(X_1, \ldots, X_m)\big)$, it can be shown (Hoeffding, 1948) that

$$\sigma_1^2 = \mathbb{V}\mathrm{ar}\big(\phi_1(X_1)\big).$$

It is evident that U-statistic is an unbiased estimator of θ. Additionally, the Law of large numbers holds: under some very weak conditions, e.g. $\mathbb{E}\big(\phi(X_1,\dots,X_m)^2\big) < \infty$ (see e.g. Hoeffding, 1948), it holds that U_n is a consistent estimator of θ, i.e. $U_n \xrightarrow{P} \theta$, $n \to \infty$.

Similarly to a U-statistic, we can define a V-statistic with kernel ϕ as

$$V_n = \frac{1}{n^m} \sum_{i_1,i_2,\dots,i_m=1}^{n} \phi(X_{i_1},\dots,X_{i_m}). \tag{1.2}$$

U- and V-statistics often arise unexpectedly as either parameter estimates or test statistics, so it is of a great importance to have insights into their asymptotic properties, especially for non-degenerate and weakly degenerate case, as those are the most commonly encountered in practice.

Two-sample U- and V-statistics

It is common to consider settings in which two IID samples $X_1,\dots,X_{n_1}$ and $Y_1,\dots,Y_{n_2}$ of d-variate random vectors are observed. Analogously to (1.1), we can define the *two-sample U-statistic* with the kernel $\phi(x_1,\dots,x_{m_1};y_1,\dots,y_{m_2})$ as

$$U_{n_1 n_2} = \frac{1}{\binom{n_1}{m_1}\binom{n_2}{m_2}} \sum_{\substack{1\leqslant i_1<i_2<\dots i_{m_1}\leqslant n_1 \\ 1\leqslant j_1<j_2<\dots j_{m_2}\leqslant n_2}} \phi\left(X_{i_1},\dots,X_{i_{m_1}};Y_{j_1},\dots,Y_{j_{m_2}}\right),$$

where the kernel is symmetric with respect to permutations of $x_1,\dots,x_{m_1}$ and $y_1,\dots,y_{m_2}$. Two-sample V-statistic is defined in the same manner.

The degeneracy of the kernel ϕ, as well as all of the other technical aspects, are defined analogously to the one-sample case. Moreover, the analogous asymptotic results in both cases hold.

Chapter 6 is devoted to the study of the energy test statistic, which is itself a two-sample V-statistic. However, since it is examined from a very specific perspective in that chapter, including the general asymptotic results for the two-sample case here would be redundant. Instead, in Section 1.4 we present only the results relevant to our work in Chapter 6, namely those by Neuhaus (1977) concerning two-sample U- and V-statistics.

For a general overview of the results in the field, the reader is referred to the books by Lehmann (1999), Koroljuk and Borovskich (2010), and Henze (2024), among others.

1.2 Asymptotic behavior of non-degenerate U- and V- statistics

Asymptotic properties of non-degenerate U-statistics will be an essential part of derivation of our results in Chapters 3 and 4, so it is useful to have the necessary results restated in this dissertation.

The following theorems, which we will refer to later, are famous results proved by Hoeffding (1948). Theorem 1.1 states the asymptotic distribution of a single non-degenerate U-statistic, and Theorem 1.2 provides the joint asymptotic distribution of two non-degenerate U-statistics calculated on the same sample. We also note that Theorem 1.1 also holds for non-degenerate V-statistic V_n, but that result will not be relevant for our further research.

Theorem 1.1 [Hoeffding (1948)]**.** *Let $X_1,\dots,X_n$ be a sample of IID d-variate random vectors and let $\phi(x_1,\dots,x_m)$ be a symmetric kernel such that $\mathbb{E}\phi(X_1,\dots,X_m) = \theta$ and $\mathbb{E}\phi^2(X_1,\dots,X_m) < \infty$.*

Suppose that $\sigma_1^2 > 0$. *Then, as* $n \to \infty$,

$$\sqrt{n}(U_n - \theta) \xrightarrow{D} \mathcal{N}(0, m^2\sigma_1^2).$$

THEOREM 1.2 [HOEFFDING (1948)]**.** *Let*

$$U_n^{(1)} = \frac{1}{\binom{n}{a}} \sum_{1 \leqslant i_1 < i_2 < \cdots < i_a \leqslant n} \phi^{(1)}(X_{i_1}, \ldots, X_{i_a})$$

and

$$U_n^{(2)} = \frac{1}{\binom{n}{b}} \sum_{1 \leqslant i_1 < i_2 < \cdots < i_b \leqslant n} \phi^{(2)}(X_{i_1}, \ldots, X_{i_b})$$

be two non-degenerate U-statistics with kernels $\phi^{(1)}$ *and* $\phi^{(2)}$, *respectively. Let* $\mathbb{E}\phi^{(1)}(X_1, \ldots, X_a)^2 < +\infty$ *and* $\mathbb{E}\phi^{(2)}(X_1, \ldots, X_b)^2 < +\infty$. *Then, as* $n \to \infty$,

$$\left(\sqrt{n}\left(U_n^{(1)} - \theta_1\right), \sqrt{n}\left(U_n^{(2)} - \theta_2\right)\right) \xrightarrow{D} \mathcal{N}(0, \Sigma),$$

where Σ *is a limit value of the covariance matrix of* $\sqrt{n}\left(U_n^{(1)} - \theta_1\right)$ *and* $\sqrt{n}\left(U_n^{(2)} - \theta_2\right)$, *which is equal to*

$$\Sigma = \begin{bmatrix} a^2\sigma_{1,(1)}^2 & ab\sigma_{11} \\ ab\sigma_{11} & b^2\sigma_{1,(2)}^2 \end{bmatrix},$$

where

$$\sigma_{11} = \mathbb{E}\left(\phi_1^{(1)}(X_1)\phi_1^{(2)}(X_1)\right) = \mathbb{C}\mathrm{ov}\left(\phi^{(1)}(X_1, X_2, \ldots, X_a),\, \phi^{(2)}(X_1, X_2', \ldots, X_b')\right),$$

where X_j' *is an independent copy of* X_j, *for every* j. *Furthermore,*

$$\sigma_{1,(1)}^2 = \mathbb{V}\mathrm{ar}\left(\phi_1^{(1)}(X_1)\right),$$

and $\sigma_{1,(2)}^2$ *is defined in a similar manner.*

1.3 Asymptotic behavior of weakly degenerate *U*- and *V*-statistics

Since Chapter 5 is devoted to gaining insight into the asymptotic distribution of a specific weakly degenerate V-statistic, this section presents several theoretical preliminaries and asymptotic results that are closely related to that topic.

Consider a sample $X_1, X_2, \ldots, X_n$ of IID d-variate random vectors with common CDF F. Let $\phi(X_1, X_2, \ldots, X_m)$ be a symmetric kernel with expected value θ, and assume that it is weakly degenerate.

Consider the weakly degenerate U-statistic with the kernel ϕ, as in (1.1). Define, for any $x, y \in \mathbb{R}^d$, the second projection

$$\phi_2(x, y) = \mathbb{E}\left(\phi(X_1, X_2, \ldots, X_m) \mid X_1 = x, X_2 = y\right) - \theta. \tag{1.3}$$

Next, define the integral operator $A : L^2(\mathbb{R}^d, \mathrm{d}F) \to L^2(\mathbb{R}^d, \mathrm{d}F)$ as

$$Ag(x) = \int_{\mathbb{R}^d} \phi_2(x, y) g(y) \, \mathrm{d}F(y). \tag{1.4}$$

It is a well-known result (e.g. Henze, 2024, pp. 119) that, under the assumption of finite second moment of the kernel, this operator is compact and self-adjoint, so its eigenvalues form a decreasing sequence of positive real numbers, that we will denote by $\{\lambda_j, j \geqslant 1\}$, with zero limit.

The following result was discovered independently by Gregory (1977) and Serfling (1980). For detailed explanation and complete proofs, see, e.g. Henze (2024), Ch. 8.

Theorem 1.3 [Gregory (1977), Serfling (1980)]**.** *Let U_n be a weakly degenerate U-statistic with the kernel ϕ as in* (1.1)*, and let $\mathbb{E}\phi^2(X_1, X_2, \dots, X_m) < \infty$. Let ϕ_2 be as in* (1.3)*, and let $\{\lambda_j, j \geqslant 1\}$ be the sequence of the eigenvalues of the associated integral operator* (1.4)*. Then, it holds that*

$$n(U_n - \theta) \xrightarrow{D} \binom{m}{2} \sum_{j=1}^{\infty} \lambda_j \big(\chi^2_{1,j} - 1\big), \quad n \to \infty, \tag{1.5}$$

where $\{\chi^2_{1,j}, j \geqslant 1\}$ are IID random variables with χ^2_1 distribution.

The corresponding result for V-statistics was discussed by both Gregory (1977) and Serfling (1980), with the main focus on the case of $m = 2$, and a kernel with zero mean. This scenario frequently arises in applications where the value of the test statistic being close to zero is indicative of the null hypothesis being true. In that case, which will also be the primary focus of Chapter 5, we consider the statistic

$$V_n = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \phi(X_i, X_j), \tag{1.6}$$

which can be decomposed as

$$\begin{aligned} V_n &= \frac{1}{n^2} \sum_{i=1}^{n} \phi(X_i, X_i) + \frac{1}{n^2} \sum_{i=1}^{n} \sum_{\substack{j=1 \\ j \neq i}}^{n} \phi(X_i, X_j) \\ &= \frac{1}{n^2} \sum_{i=1}^{n} \phi(X_i, X_i) + \frac{n-1}{n} \frac{1}{\binom{n}{2}} \sum_{1 \leqslant i < j \leqslant n} \phi(X_i, X_j). \end{aligned}$$

So, it follows that

$$n V_n = \frac{1}{n} \sum_{i=1}^{n} \phi(X_i, X_j) + n U_n - U_n. \tag{1.7}$$

Having this, we immediately obtain the following theorem.

Theorem 1.4. *Let V_n be a weakly degenerate V-statistic as in* (1.6)*, and let $\mathbb{E}|\phi(X_1, X_1)| < \infty$, $\mathbb{E}\phi^2(X_1, X_2) < \infty$, and $\mathbb{E}\phi(X_1, X_2) = 0$. Then, as $n \to \infty$, it holds that*

$$n V_n \xrightarrow{D} \mathbb{E}\phi(X_1, X_1) + \sum_{j=1}^{\infty} \lambda_j \big(\chi^2_{1,j} - 1\big),$$

where $\{\chi^2_{1,j}, j \geqslant 1\}$ are IID random variables with χ^2_1 distribution, and $\{\lambda_j, j \geqslant 1\}$ are eigenvelues of the integral operator defined in (1.4)*.*

We say that an operator is trace-class if its singular values are summable. In the case of compact and self-adjoint operator this condition is equivalent to $\sum_{j=1}^{n} |\lambda_j| < \infty$, where λ_j are

its eigenvalues. If we additionally assume that the operator A defined in (1.4) is trace-class, then it follows (see e.g. Brislawn, 1991) that

$$\mathbb{E}\phi(X_1, X_1) = \mathbb{T}\mathrm{r}(A) = \int_{\mathbb{R}^d} \phi(x, x)\mathrm{d}x = \sum_{j=1}^{\infty} \lambda_j.$$

As a direct consequence, (1.7) becomes

$$nV_n \xrightarrow{D} \sum_{j=1}^{\infty} \lambda_j \chi^2_{1,j}, \quad n \to \infty.$$

Determining whether integral operator (1.4) is trace-class is not a trivial task, and characterization-based criteria tend to be very abstract. Notable results in this area include the criteria established by Brislawn (1988, 1991). Additionally, Simon (2005) provides valuable insights and further references on this topic.

1.4 Results of Neuhaus (1977) for weakly degenerate two-sample *U*- and *V*-statistics

In this section, we present some of the results obtained by Neuhaus (1977) which are related to our work in Chapter 6. We note that this is only part of those results, and that they are slightly more general than those presented here. Moreover, we heavily modified the notation from the original paper to best suit our needs.

Consider two independent samples $X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots Y_m$ of IID d-variate random vectors with common CDF F, and let $h(x_1, x_2; y_1, y_2)$ be a measurable kernel symmetric with respect to the mutual permutations of x_1 and x_2, or y_1 and y_2, and let $\mathbb{E}\big(h^2(X_1, X_2; Y_1, Y_2)\big) < \infty$. Additionally, assume that h is degenerate kernel in a way that

$$h_{1,1}(x, y) := \mathbb{E}\Big(h(X_1, X_2; Y_1, Y_2) \Big| X_1 = x,\ Y_1 = y \Big) = 0$$

for almost every pair (x, y), under the distribution of (X_1, Y_1).

According to the Spectral Theorem (for a modern reference see, e.g. Henze, 2024, Th. 8.14), there exists an orthonormal sequence $f_1, f_2, \ldots \in L^2\big(\mathbb{R}^{2d}, \mathrm{d}F(x)\mathrm{d}F(y)\big)$ of functions such that $\int f_j(x, y)\mathrm{d}F(x)\mathrm{d}F(y) = 0$ for $j \geqslant 1$, and a diminishing sequence $\lambda_1, \lambda_2, \ldots$ of positive real numbers such that $\sum_{j=1}^{\infty} \lambda_j^2 = \mathbb{E}\big(h^2(X_1, X_2; Y_1, Y_2)\big) < \infty$, so that, if

$$h^s(x_1, x_2; y_1, y_2) = \sum_{j=1}^{s} \lambda_j f_j(x_1, y_1) f_j(x_2, y_2),$$

then

$$\lim_{s \to \infty} \mathbb{E}\big[(h(X_1, X_2; Y_1, Y_2) - h^s(X_1, X_2; Y_1, Y_2))^2 \big] = 0.$$

Moreover, $\lambda_1, \lambda_2, \ldots$ are eigenvalues of the integral operator

$$A : L^2\big(\mathbb{R}^{2d}, \mathrm{d}F(x)\mathrm{d}F(y)\big) \to L^2\big(\mathbb{R}^{2d}, \mathrm{d}F(x)\mathrm{d}F(y)\big)$$

defined as

$$Ag(x_1, y_1) = \int_{\mathbb{R}^{2d}} k(x_1, x_2; y_1, y_2) g(x_2, y_2)\mathrm{d}F(x_2)\mathrm{d}F(y_2),$$

with orthonormal eigenfunctions $f_1, f_2, \ldots$ and

$$k(x_1, x_2; y_1, y_2) = h(x_1, x_2; y_1, y_2) - \mathbb{E}[h(X_1, X_2; Y_1, Y_2)].$$

Let W_{1j}, W_{2j}, $j \geqslant 1$, be independent Brownian motions on $[0,1]$. For any $0 \leqslant t_1, t_2 \leqslant 1$, let

$$U_{nm}(t_1, t_2) = \frac{1}{nm(n+m)} \sum_{i=1}^{\lceil nt_1 \rceil} \sum_{\substack{j=1 \\ j \neq i}}^{\lceil nt_1 \rceil} \sum_{k=1}^{\lceil mt_2 \rceil} \sum_{\substack{l=1 \\ l \neq k}}^{\lceil mt_2 \rceil} h\big(X_i, X_j; Y_k, Y_l\big), \tag{1.8}$$

where $\lceil \cdot \rceil$ is the ceiling function, and let

$$U(t_1, t_2) = \sum_{j=1}^{\infty} \lambda_j \Big[\big(a_j t_2 W_{1j}(t_1) + b_j t_1 W_{2j}(t_2)\big)^2 - \big(a_j^2 t_2 + b_j^2 t_1\big) t_1 t_2 \Big], \tag{1.9}$$

where

$$a_j^2 = \beta \int \big(\mathbb{E} f_j(x, Y_1)\big)^2 \, \mathrm{d}F(x), \quad b_j^2 = \alpha \int \big(\mathbb{E} f_j(X_1, y)\big)^2 \, \mathrm{d}F(x), \quad j \geqslant 1, \tag{1.10}$$

and α and β such that $n/(n+m) \to \alpha$ and $m/(n+m) \to \beta$. We observe U_{nm} as a random element in the space $D_2 := D([0,1]^2)$ of all real functions on $[0,1]^2$ with no discontinuities of the second kind, equipped with the Skorohod metric (see e.g. Billingsley, 1968, Sec. 14.2).

The following result is due to Neuhaus (1977).

THEOREM 1.5 [NEUHAUS (1977)]**.** *As $n, m \to \infty$, $n/(n+m) \to \alpha$, $m/(n+m) \to \beta$, it holds that*

$$U_{nm} \xrightarrow{D} U, \quad \text{in } D_2. \tag{1.11}$$

1.5 Results of De Wet and Randles for parameter-dependent kernels

Sections 1.2 and 1.3 provided an overview of the possible asymptotic distributions of U- and V-statistics in the non-degenerate and weakly degenerate cases, respectively. However, it is common in the hypothesis testing literature for the kernel of a statistic to depend on unknown parameters of the underlying distribution. In such cases, the statistic cannot be directly computed, as the true parameter values are not available. A standard approach is to estimate the unknown parameters and then proceed with the computation using these estimates. This naturally raises the question of whether, and in what way, parameter estimation affects the asymptotic distribution of the statistic. More specifically, it is of special interest to identify which estimators, when substituted for the true parameters, leave the asymptotic distribution unchanged. First results addressing this question in the non-degenerate case were provided by Randles (1982), and were later followed by those for weakly degenerate case by De Wet and Randles (1987). The latter were generalized by Cuparić et al. (2022).

1.5.1 Non-degenerate case

The conditions under which the asymptotic distribution of a non-degenerate U-statistic with estimated parameters remains unchanged were first established by Randles (1982). Randles also presented analogous results for V-statistics. Due to not being directly related to our results in Chapter 4, they are omitted here. For details, we refer to the original paper.

Consider a random sample $X_1,\dots,X_n$ of d-variate random vectors and let $h(X_1,\dots,X_m;\gamma)$ be a symmetric kernel with expected value $\theta(\gamma)=\mathbb{E}_\lambda h(X_1,\dots,X_m;\gamma)$, where λ denotes the true value of the parameter, and the expectation is taken with respect to this value. To estimate the expected value of the kernel, consider the U-statistic

$$U_n(\gamma)=\frac{1}{\binom{n}{m}}\sum_{1\leqslant i_1<i_2<\cdots<i_m\leqslant n} h(X_{i_1},\dots,X_{i_m};\gamma) \tag{1.12}$$

and assume that it is non-degenerate for every γ. Since the true parameter value λ is commonly unknown, it is usually estimated from the data by some estimator $\hat{\lambda}$. Consequently, the statistic $U_n(\lambda)$ is replaced by the plug-in version $U_n(\hat{\lambda})$.

For any fixed value of the parameter, Theorem 1.1 establishes that the statistic is asymptotically normally distributed. Randles (1982) provided the set of conditions, presented below, under which the asymptotic distributions of $\sqrt{n}(U_n(\lambda)-\theta(\lambda))$ and $\sqrt{n}\big(U_n(\hat{\lambda})-\theta(\lambda)\big)$ coincide.

1. [Orig. *Condition 2.2*] Suppose that

$$\sqrt{n}\big(\hat{\lambda}-\lambda\big)=O_{\mathbb{P}}(1),\quad n\to\infty. \tag{1.13}$$

2. [Orig. *Condition 2.3*] Let $D(\gamma,d)$ be a ball centered at γ with radius d. Suppose there exists a neighborhood $K(\lambda)$ around λ and a positive constant K_1 such that if $\gamma\in K(\lambda)$ and $D(\gamma,d)\subseteq K(\lambda)$, then

$$\mathbb{E}\left(\sup_{\gamma'\in D(\gamma,d)}\left|h(X_1,\dots,X_m;\gamma')-h(X_1,\dots,X_m;\gamma)\right|\right)\leqslant K_1 d \tag{1.14}$$

and

$$\lim_{d\to 0}\mathbb{E}\left(\sup_{\gamma'\in D(\gamma,d)}\left|h(X_1,\dots,X_m;\gamma')-h(X_1,\dots,X_m;\gamma)\right|^2\right)=0. \tag{1.15}$$

Remark 1.1. Under the condition that $|h(x_1,\dots,x_m;\gamma')-h(x_1,\dots,x_m;\lambda)|<M_1$ for some positive M_1, every $x_1,\dots,x_m$ and every γ in a neighborhood of λ, it holds that (1.14) implies (1.15) (Randles, 1982, Lemma 2.6).

3. [Orig. *Condition 2.9A*] Let (1.13) hold, and let $\sqrt{n}(U_n(\lambda)-\theta(\lambda))\xrightarrow{D}\mathcal{N}(0,m^2\sigma_1^2)$, where σ_1^2 is the covariance seen in Theorem 1.1. Assume that $\theta(\gamma)$ has zero differential at $\gamma=\lambda$.

4. [Orig. *Condition 2.9B*] Assume that $\theta(\gamma)$ has nonzero differential at $\gamma=\lambda$, and additionally assume that

$$\sqrt{n}\big(U_n(\lambda)-\theta(\lambda),(\hat{\lambda}-\lambda)^T\big)\xrightarrow{D}\mathcal{N}_{d+1}(0,\Sigma).$$

Having the conditions stated, we can now formulate the result by Randles (1982).

Theorem 1.6 [Randles (1982)]**.** *If Condition 2.3 holds alongside with one of the Conditions 2.9A and 2.9B, then*

$$\sqrt{n}\big(U_n(\hat{\lambda})-\theta(\lambda)\big)\xrightarrow{D}\mathcal{N}(0,m^2\sigma_1^2),\quad n\to\infty,$$

where σ_1^2 is as in Theorem 1.1.

1.5.2 Weakly degenerate case

Since Chapter 5 is devoted to gaining insight into the asymptotic distribution of weakly-degenerate V-statistics with estimated parameters under certain missing data scenarios, we first present the necessary results for the complete-sample setting. Given the scope of our work, we focus exclusively on the key results for V-statistics. For a more detailed analysis, including the analogous results for U-statistics, we refer to the original paper by De Wet and Randles (1987).

Note that the notation used in this section differs slightly from that of the original paper, mostly for cosmetic reasons, and has been adjusted to align with the notation from Baringhaus and Henze (1988), which we build upon in Chapter 5.

Let $X_1, \ldots, X_n$ be IID d-variate random vectors, and let us consider the weakly degenerate V-statistic

$$V_n(\lambda) = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} h(X_i, X_j; \lambda),$$

whose symmetric kernel depends on the unknown parameter λ, and has zero mean for every value of the parameter. Let $\hat{\lambda}$ be a consistent estimator of λ based on $X_1, \ldots, X_n$. Then it is natural to use

$$V_n(\hat{\lambda}) = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} h(X_i, X_j; \hat{\lambda}).$$

Additionally, we assume that the kernel h admits the representation

$$h(x_1, x_2; \lambda) = \int_{\mathbb{R}^d} g(x_1, t; \lambda) g(x_2, t; \lambda) \, \mathrm{d}M(t),$$

for some function g and a finite positive measure M on $\mathbb{R}^d$.

The conditions of De Wet and Randles (1987) are as follows.

1. [Orig. *Condition 2.9*] Suppose

   $$\epsilon(t; \gamma) = \mathbb{E}_\lambda \left[g(X_1, t; \gamma) \right]$$

   exists and $\epsilon(t; \lambda) = 0$ for every t and every γ in some neighborhood of λ. Additionally, assume $\varepsilon(t; \gamma)$ is L^2-differentiable at $\gamma = \lambda$, and let $\epsilon_1(t; \lambda)$ be the differential.

2. [Orig. *Condition 2.10*] Suppose

   $$\hat{\lambda} = \lambda + \frac{1}{n} \sum_{i=1}^{n} \alpha(X_i) + o_{\mathbb{P}}\left(\frac{1}{\sqrt{n}} \right),$$

   where $\mathbb{E}[\alpha(X_i)_r] = 0$ and $\mathbb{E}[\alpha(X_i)_r \alpha(X_i)_{r'}]$ is finite for all $1 \leqslant r < r' \leqslant d$ (here subscript denotes the component of the vector).

3. [Orig. *Condition 2.11*] Suppose that there exists $M^* > 0$ and a neighborhood $K(\lambda)$ of λ such that

   (a) if $\gamma \in K(\lambda)$ and $D(\gamma, r)$ is a ball centered in γ with radius r such that $D(\gamma, r) \subseteq K(\lambda)$, then

   $$\int_{-\infty}^{\infty} \left(\mathbb{E}\left[\sup_{\gamma' \in D(\gamma, r)} \left| g\left(X_i, t; \gamma'\right) - g\left(X_i, t; \gamma\right) \right| \right] \right)^2 \mathrm{d}M(t) \leqslant M^* r^2,$$

   and

(b) for any $\varepsilon > 0$ there exists a $r^* > 0$ such that $0 < r < r^*, \gamma \in K(\lambda)$ and $D(\gamma, r) \subset K(\lambda)$ imply

$$\int_{-\infty}^{\infty} \mathbb{E}\left[\sup_{\gamma' \in D(\gamma,r)} \left|g\left(X_i, t; \gamma'\right) - g\left(X_i, t; \gamma\right)\right|^4\right] \mathrm{d}M(t) < \varepsilon.$$

Let

$$V_n = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} h_*\left(X_i, X_j\right),$$

where

$$h_*(x, y) = \int_{-\infty}^{\infty} \left[g(x, t; \lambda) + \epsilon_1(t; \lambda)' \alpha(x)\right] \cdot \left[g(y, t; \lambda) + \epsilon_1(t; \lambda)' \alpha(y)\right] \mathrm{d}M(t).$$

THEOREM 1.7 [DE WET, RANDLES (1987)]**.** *Let $X_1, \ldots, X_n$ be IID d-variate random vectors with CDF $F(x)$. Suppose Conditions 2.9, 2.10 and 2.11 hold, and that*

$$\mathbb{E}\left[h_*^2(X_1, X_2)\right] < \infty \quad \text{and} \quad \mathbb{E}[h_*(X_1, X_1)] < \infty.$$

Let $\{\delta_k\}$ denote the eigenvalues of the integral operator A defined by

$$Aq(x) = \int_{-\infty}^{\infty} h_*(x, y) q(y) dF(y).$$

Then, as $n \to \infty$,

$$n\left(V_n(\hat{\lambda}) - V_n\right) \xrightarrow{P} 0$$

and

$$n V_n(\hat{\lambda}) \xrightarrow{D} \sum_{k=1}^{\infty} \delta_k \chi_{1,k}^2$$

where $\chi_{1,k}^2, k = 1, 2, \ldots$, are independent χ_1^2 variates.

The theorem states that the effect of the estimation of λ is entirely captured by ϵ_1. It establishes that $n V_n(\hat{\lambda})$ and $n V_n$ have the same asymptotic distribution. However, if ϵ_1 is equal to zero, as in the case studied by Baringhaus and Henze (1988), then $V_n = V_n(\lambda)$. In other words, in that scenario $n V_n(\hat{\lambda})$ has the same asymptotic distribution as $n V_n(\lambda)$.

Chapter 2

Mathematical framework for missing data

Handling missing data is a critical aspect of statistical analysis, as its presence can significantly affect the validity and reliability of the conclusions. Failure to properly manage missing data may lead to biased estimates, distorted results, and reduced accuracy of statistical inference. However, many applied research studies either insufficiently describe their treatment of missing data or do not acknowledge its presence entirely. A useful overview of this issue for survey data is provided by Mirzaei et al. (2022).

The main goal of this chapter is to present the key definitions and theoretical concepts of missing data analysis that are essential for understanding the subsequent chapters. It is not intended as a comprehensive monograph on the topic, that is, we do not attempt to provide an exhaustive overview of the theoretical foundations or detailed guidelines on the practical application of various methods. These topics are thoroughly addressed in the existing literature. For readers interested in the theoretical aspects, we recommend the monograph by McKnight et al. (2007) as a gentle introduction, particularly suitable for those with a more elementary background in mathematics and statistics. For a more advanced and comprehensive treatment, the well-known monograph by Little and Rubin (2019) provides deeper insight and greater mathematical rigor. For readers interested in the practical application of missing data handling methods, we recommend the monograph by Enders (2022) as a general resource, and the monograph by Van Buuren (2018) for a focused treatment of imputation techniques. The latter also includes many sections that delve into the theoretical and mathematical foundations of the methods, but it is structured in a way that allows readers to skip these parts without losing the ability to follow the rest of the text. For a more concise summary of recent developments in the field, see the paper by Enders (2023).

Section 2.1 provides a brief overview of the development of missing data analysis as a distinct statistical discipline. In Section 2.2, we introduce the definitions of the three main types of missingness mechanisms, MCAR, MAR, and MNAR, along with the underlying rationale. Given that these definitions were not always clear throughout the historical evolution of the field, Section 2.3 is dedicated to addressing and clarifying that ambiguity. Finally, since Chapter 3 introduces a novel test for MCAR, Section 2.4 presents a historical overview of the development of various MCAR testing procedures.

2.1 Origins

The development of missing data analysis as a distinct statistical discipline coincided with the increasing complexity of real-world datasets and the growing difficulty of collecting high-quality complete data. Although researchers in the early to mid-20th century frequently encountered incomplete datasets, they often relied on relatively simple methods to address the issue. The most commonly used approach was complete-case analysis, which involves discard-

ing all observations with at least one missing value. An alternative strategy involved filling in the missing values using straightforward methods, a practice now known as *imputation*. The most commonly used imputation techniques included replacing missing values with the sample mean or median. However, as is now well understood, such naive techniques can lead to severely biased estimates and undermine the validity of statistical inference, particularly when the missingness mechanism is strongly dependent on the data, or when missingness rates are high (see e.g. Aleksić and Milošević, 2025a, for a recent reference). Nevertheless, when the missingness is random and occurs at moderate rates, these simplistic techniques often perform adequately in practice. This observation underscored the importance of formally defining the concepts of *random missingness* and *mild or moderate missingness rates*, as a foundation for justifying the use of such methods.

A major turning point in the development of missing data analysis was the publication of the seminal paper *Inference and Missing Data* by Rubin (1976). In this work, Rubin introduced a formal probabilistic framework for handling missing data, therefore laying the foundation for the modern theory of missing data analysis. One of the key contributions of the paper was the formal definition of the concept of *missing at random (MAR)*, which will be discussed in detail in Section 2.2. Rubin's primary focus was on parametric inference, both frequentist and Bayesian, and he investigated the conditions under which the missingness mechanism could be ignored in the analysis.

The *Expectation–Maximization Algorithm (EM Algorithm)*, proposed shortly thereafter by Dempster et al. (1977), provided a practical method for computing maximum likelihood estimates (MLEs) from incomplete datasets, and it remains widely used today.

A very common approach when working with missing data is to impute the missing values using some imputation method. In practice, after the imputation, the analysis is conducted as if the data were complete. However, the imputation clearly changes the distribution of the data; a clear example will be presented in Chapter 4.

Another important contribution by Rubin is the development of multiple imputation (Rubin, 1987), a method in which missing values are imputed multiple times to create several complete datasets. Each dataset is analyzed separately, and the resulting estimates are then combined in a certain way, allowing for the incorporation of uncertainty due to missingness.

The book *Statistical Analysis with Missing Data* by Little and Rubin (1987) quickly became one of the most influential references in the field. It continues to serve as a foundational text, now in its third edition(Little and Rubin, 2019), and remains widely used by both researchers and practitioners. We highly recommend it as a comprehensive resource on both the theoretical and applied aspects of missing data analysis.

2.2 Missingness mechanisms

Suppose that we have a sample $X_1, \ldots, X_n$ of IID d-variate random vectors, i.e. each X_j, for $j = 1, 2, \ldots, n$, can be represented as $X_j = \left(X_j^{(1)}, \ldots, X_j^{(d)}\right)$, where we mainly consider the components of X_j to be real-valued random variables, although, as we will see, the definitions also hold for more general spaces. Some of the components of X_j may be missing. For every X_j, we introduce the *response indicator vector* $R_j = \left(R_j^{(1)}, \ldots, R_j^{(d)}\right)$, where

$$R_j^{(k)} = \begin{cases} 1, & \text{if } X_j^{(k)} \text{ is observed,} \\ 0, & \text{if } X_j^{(k)} \text{ is missing,} \end{cases} \qquad k = 1, 2, \ldots, d.$$

The term *response indicator* originated in Rubin's early work on survey analysis and has since been adopted in various other data analysis contexts, including settings such as measurement data, where no actual respondents are involved. We also note that the notation in which the

index of a sample element appears as a subscript and the index of a vector component as a bracketed superscript is convenient when first introducing the concept. However, alternative notational conventions are commonly used throughout this thesis and in the broader literature, including forms such R_{jk}, $R_{j,k}$, R_j^k, R_j^X, and many others. Additionally, *missingness indicators* are sometimes used instead, which are equal to 1 if the value is missing, and zero if it is observed.

In any real-world data analysis, the analyst works with the realized sample $x_1, \ldots, x_n$, and the corresponding realizations $r_1, \ldots, r_n$ of the response indicators. The latter, which take values in $\{0,1\}^d$, are referred to as *response patterns* (or *missingness patterns*). The observed components of a vector x, with respect to a given response pattern r, that is, the subvector of x consisting of components corresponding to the entries of r that are equal to 1, will be denoted by x_{obs}. Missing elements of x will be denoted by x_{mis}. Similarly, for a random vector X, we have X_{obs} and X_{mis}.

To summarize, we have a sample of n IID realizations of a d-variate random vector X (or, more generally, a random element from the Cartesian product of d different spaces), along with the corresponding response indicator vector R, which takes values, referred to as response (or missingness) patterns, in $\{0,1\}^d$. Observed (with respect to R) elements of X are denoted by X_{obs}, and those that are missing by X_{mis}. The probability distribution of R is known as the *missingness mechanism*. This distribution may or may not depend on the data itself, and this difference leads to the definitions of the three main types of missingness mechanisms. The following definitions, adjusted for modern notation, are due to Rubin (1976) and Little and Rubin (1987).

We say that the data are *missing completely at random (MCAR)*, if the missingness mechanism does not depend on the data, neither observed nor missing, i.e.

$$\mathbb{P}\big(R = r \mid X_{\text{obs}}, X_{\text{mis}}\big) = \mathbb{P}\big(R = r\big). \tag{2.1}$$

Under the MCAR assumption, the missingness mechanism depends only on its own distributional parameters. The data are *missing at random (MAR)*, if the missingness mechanism is conditionally independent of the missing values, given the observed values, that is, it may depend on the observed data but not on the missing data. Formally,

$$\mathbb{P}\big(R = r \mid X_{\text{obs}}, X_{\text{mis}}\big) = \mathbb{P}\big(R = r \mid X_{\text{obs}}\big). \tag{2.2}$$

Finally, if the missingness mechanism depends on both the observed and the missing data, we say that the data are *missing not at random (MNAR).*

As briefly discussed in Section 2.1, the MCAR assumption is highly desirable in practice, as it allows for the use of simple missing data handling methods without compromising the validity of the analysis. In many settings of statistical inference, particularly parametric ones, the MAR assumption is often sufficient. For instance, likelihood-based estimation and Bayesian inference can still yield valid results under MAR. Seaman et al. (2013) provide a comprehensive overview of various approaches to statistical inference with incomplete data, along with the assumptions required for their validity. For a detailed treatment of Bayesian inference with incomplete data, we refer the reader to the Chapter 18 of the monograph by Gelman et al. (2014).

Since this thesis focuses on nonparametric inference, a detailed presentation of those results is omitted; a brief discussion follows in Section 2.3.

2.3 Ambiguity throughout the literature

There are several points of ambiguity regarding the definitions of MCAR, MAR, and MNAR presented in Section 2.2 which are important to address.

One such issue is whether equations (2.1) and (2.2) must hold for *every* $r \in \{0,1\}^d$, or only for the specific response pattern observed in a given realized sample. In this thesis, we adopt the stronger requirement: the condition must hold for all possible r. This formulation is standard in frequentist inference, both parametric and nonparametric, where conclusions are based on the idea of repeated sampling. On the other hand, when dealing with, e.g., likelihood-based or Bayesian inference, only the realized sample is of importance.

Little and Rubin (2019), being primarily interested in those types of inference, define the data to be MCAR (and similarly MAR, or MNAR) if (2.1) holds for realized response pattern. If (2.1) holds for every r, then the data are said to be *missing always completely at random (MACAR)*. Analogously, one can define the concepts of *missing always at random (MAAR)* and *missing not always at random (MNAAR)*. Note, however, that under frequentist inference, the notion of *missing always not at random (MANAR)* is not particularly meaningful: for the distribution of R to depend on X, it suffices that there exists even a single r for which the conditional probability $\mathbb{P}\big(R = r \mid X_{\text{obs}}, X_{\text{mis}}\big)$ cannot be simplified any further. In some other papers (e.g. Seaman et al., 2013), the term *realized MCAR (RMCAR)* is used for the definition of MCAR presented in Section 2.2, while *everywhere MCAR (EMCAR)* is used in place of MACAR. Analogously, the terms RMAR and EMAR are used for the MAR setting. Additionally, Gelman et al. (2014) treats censored data as a separate category from MNAR, and Potthoff et al. (2006) defined the *MAR+* missingness, aiming to define class of alternatives to MCAR that can be detected by their proposed test.

Another important concern regarding the definitions given in Section 2.2 arises when the sample $X_1, \ldots, X_n$ does not consist of IID elements. This situation occurs, for example, in time series analysis. In such cases, for each time point t, we observe a data point $X(t)$ and a corresponding response indicator $R(t)$. There are at least two natural ways to define MCAR in this context: one possibility is that $R(t)$ is independent of $X(t)$ for every t; another is that $R(s)$ is independent of $X(s)$ for all $s \leqslant t$. Additionally, one can view the time series as a random function taking values in a suitable metric or Hilbert space, and define MCAR by requiring the independence of the random elements X and R. Different definitions are better suited for different real-world scenarios, depending on the structure and goals of the analysis.

Going further, the notation X_{obs} and X_{mis} is also ambiguous: observed and missing according to what? In fact, the existence of a response pattern that determines the division of data into observed and missing components is implicitly assumed. A more precise notation for the observed (and, analogously, missing) parts of the data, conditional on a given response pattern r, would be X^r_{obs}, $X_{\text{obs}}(r)$, $X_{\text{obs},r}$, or similar. Seaman et al. (2013) use $o(X, r)$ and $\overline{o}(X, r)$ to denote the observed and missing components of X under the response pattern r. Although mathematically rigorous, this notation is cumbersome in long calculations and is rarely used.

Some authors, particularly in more applied and less mathematically rigorous fields, define MCAR informally with statements such as: *"The data are MCAR if the missingness does not depend on either the observed or the missing data."* While this aligns with the formal definitions given in Section 2.2, it can also be interpreted as referring to the entire dataset, in which case we have the *response indicator matrix*. However, in that context, it is unclear what kinds of precise mathematical objects are X_{obs} and X_{mis}.

In many contexts, but especially in the case of likelihood-based inference, it is important to emphasize the distinction of parameters. To be more specific, the missingness mechanism, i.e. the probability distribution of R depends on its own distributional parameters, say, η, and the distribution of X on parameters θ. One of the key assumptions for validity of statistical inference in those settings is that those two distributions do not share any parameters, i.e. that η and θ are distinct (see e.g. Seaman et al., 2013, Sec. 5).

As seen in the previous discussion, different contexts may require slightly different, though closely related, definitions of MCAR, MAR, and MNAR. For this reason, it is recommended to

explicitly state the formal definition of the missingness mechanisms being used whenever presenting work that involves missing data. Fortunately, this practice has largely become standard in the modern statistical literature.

We are certainly not the first to raise concerns regarding the ambiguity in the definitions of missingness mechanisms. Several papers have been devoted entirely to clarifying these issues, the most widely cited being the work by Seaman et al. (2013). Other valuable contributions that we sincerely recommend include those by Mealli and Rubin (2015, 2016), Doretti et al. (2018), and the series of papers by Galati (2018a,b,c, 2019).

2.4 Historical overview of MCAR tests

There has been quite a lot of interest in testing the MCAR assumption. First results were developed in the 1980's: for categorical data by Fuchs (1982), and for the Gaussian data by Little (1988). As far as we know, Little's MCAR test, that is based on comparing the MLEs across missingness patterns, remains the most widely used test in practice. However, as we will see in Chapter 3, it is highly sensitive to the assumption of data being sampled from the multivariate normal distribution.

Diggle (1989) considered missing data in the context of repeated measurements, i.e. when a time-ordered sequence of measurements is made on some participants in an experiment. He considered a special case of missingness called a *dropout*, where a sequence of measurements is terminated prematurely, and developed the class of procedures that test whether dropouts in the data occur randomly, in the sense that they are not related to any of the past measurements. The methodology involves selecting a score function, where large values indicate rejection of the null hypothesis, and applying the normal approximation when feasible. Ridout and Diggle (1991) presented some improvements in terms of flexibility, utilizing logistic regression. Park and Davis (1993) extended Little (1988) test to incomplete repeated categorical data. Similarly, Park et al. (1993) relied on Little's test to make a MCAR test for repeated measurements. Following the idea of Park and Davis, Park and Lee (1997) constructed a MCAR test for the incomplete longitudinal data in the framework of generalized estimating equations. Listing and Schlittgen (1998) developed a test for random dropouts in clinical trials by comparing the means of the individuals that stay, and those that drop out.

Another test for the framework of generalized estimating equations, but for independent observations, came from Chen and Little (1999), and generalized the idea by Little (1988). Qu and Song (2002) proposed a more unified generalized score-type test for ignorable missingness in longitudinal data.

Kim and Bentler (2002) studied tests based on weighted generalized least squares methods, and compared them to the likelihood-based tests, such as Little's test, in terms of type I error and power behavior in small sample sizes. The comparison examines the homogeneity of means and covariance matrices across missing data patterns.

A further logistic regression-based testing procedure for MCAR tailored for medical longitudinal data was developed by Fairclough (2002). The main idea of the procedure is to study the dependence between the response indicators and the quality of life scores.

Although testing MCAR vs. MAR is not possible in the general case, since the data needed for that testing are missing, Potthoff et al. (2006) proposed the test for *MAR+* assumption.

The idea of testing MCAR by comparing the covariance matrices across the missing data patterns came with Jamshidian and Schott (2007). Jamshidian and Mata (2008) constructed a test for distinguishing MCAR from MNAR by noting that the maximum-likelihood estimates across random data subsamples will have the same asymptotic distribution under MCAR, but not under MNAR.

Fielding et al. (2009) presented a real-data empirical comparison of four tests in the context of quality of life outcomes. Specifically, the tests of Little (1988), Listing and Schlittgen (1998), Ridout and Diggle (1991), and Fairclough (2002) were applied to several datasets to assess differences in inferential results.

Jamshidian and Jalal (2010) have considered a test for MCAR that relies on imputing the dataset and then conducting the complete-data procedures. The data are grouped by missingness patterns, and the variances across groups of data are then compared. Jamshidian and Yuan (2013) improved the results of Jamshidian and Mata (2008) by approximating the asymptotic distribution rather than using the bootstrap method.

Lin (2013) developed a probability based framework for testing MCAR that appeared comparable to Little's MCAR test in terms of power for a large number of studied scenarios.

Jamshidian and Yuan (2014) gave an overview of then available MCAR test that are based on either homogeneity of parameters or homogeneity of distributions across the missingness patterns. Additionally, they proposed a novel nonparametric test for MCAR that is based on pairwise comparison of marginal distributions of the data, considering one variable at a time fixed. Li and Yu (2015) also considered a nonparametric test for MCAR. The procedure first splits all of the data into categories by missingness patterns, and then uses Rizo-Székely dissimilarity measure to compare distributions across patterns. The bootstrap algorithm is utilized afterwards to approximate the p-value of the test.

Yuan et al. (2018) showed that, under normality, MLEs for different missingness patterns can converge to the same values, possibly not the true ones, even under MAR or MNAR. As a result, tests for MCAR based on comparing means and covariances across patterns cannot be safely used.

Zhang et al. (2019) noted that most MCAR tests do not offer a method for a subsequent estimation once MCAR is rejected, and they presented a unified likelihood approach for both MCAR testing and subsequent estimation that appeared to behave well in the observed (although limited) scenarios.

Bojinov et al. (2020) considered testing MAAR, where response mechanism does not depend on the data not only for the observed, but for any possible missingness pattern. They note that under certain regularity conditions, MAAR can be tested from the observed data only, and propose three diagnostic procedures that rely on testing the dependence between response indicators and fully observed variables.

Spohn et al. (2021) introduced the test that measures distributional differences across missing data patterns using Kullback-Leibler divergence. Rouzinov and Berchtold (2022) tested MCAR by fitting the linear regression model on the complete cases, and then comparing distributional differences of predicted values for missing and observed data.

For the case of hidden Markov models, Chassan and Concordet (2023) developed a MCAR which does not require grouping the data by patterns, but are based on the estimates of conditional (given the latent state of the Markov chain) probabilities of missingness.

Lately, the measure of *compatibility* was utilized by Berrett and Samworth (2023) for testing MCAR. Their key point is that there can be no test that can reject MCAR if the class of marginal distributions is compatible, i.e. they were successful in describing the exact class of non-detectable alternatives to MCAR. They related the concept of compatibility testing to MCAR testing in the discrete case. Bordino and Berrett (2024) compared compatibility of covariance matrices across missing data patterns to construct a MCAR test for the incomplete data that do not need to be discrete. The formal definition of compatibility can be found in either of these works, and can be traced back to Sklar (1959) and the theory of copulas.

Dealing with functional data, we refer the reader to the recent test by Ofner et al. (2025).

Most of the existing statistical tests for MCAR are based on comparing some measure across different missing data patterns. To the best of our knowledge, there were no tests constructed

using the rationale of checking the linear dependence between the response indicators and fully observed data columns. This changed with the tests that are the subject of Chapter 3 and have been presented in the studies of Aleksić (2024, 2025a).

Chapter 3

Testing the MCAR assumption utilizing the properties of U-statistics

In this chapter, we present our novel test for assessing the MCAR assumption. The test builds upon the theory of non-degenerate U-statistics to establish its asymptotic properties. It was first introduced by Aleksić (2024) and generalized a year later by Aleksić (2025a). This chapter extends this generalization by providing a deeper examination of the null distribution of the test statistic and offering further insights into potential extensions of the proposed methodology. Furthermore, a class of detectable alternatives is described in more detail.

The original version of the test, which we will refer to as *the old, first*, or sometimes *the original test*, is introduced in Section 3.1 and its asymptotic null distribution is derived. In Section 3.2, it is shown that the proposed test statistic coincides with the well-known Little's statistic in the case of univariate nonresponse (Little, 1988). The generalization of the test, which we will refer to as *the novel*, or *the new test* is introduced in Section 3.3, along with the derivation of its asymptotic properties. Extensive simulation study is conducted in Section 3.4, where the old test, the novel test and Little's MCAR test are compared in terms of the preservation of type I error and the power performance.

Additional simulation results supporting the findings presented in this dissertation can be found in the supplementary materials of the two aforementioned papers, as well as on the author's `GitHub` page (Aleksić, 2025b).

3.1 Main idea and first version of the test

The objective of this section is to introduce the original variant of our test statistic and to establish its asymptotic properties. We proceed step by step, ensuring that the rationale behind the test is presented in a clear and structured manner. We begin with the case of two-dimensional data with univariate nonresponse, which allows us to illustrate the main ideas in a simple setting, and then gradually extend the discussion to the general multivariate case. The ultimate goal of this section is to present Theorem 3.2, which formalizes the asymptotic properties of the proposed statistic and lays the foundation for the subsequent analysis.

3.1.1 Two-dimensional data with univariate nonresponse

Suppose we have data from a bivariate distribution that can be represented by the random vector (X, Y), where $\mathbb{E}(X^2) < \infty$, and let us model our sample as its n IID copies, which we

expand to obtain

$$\begin{bmatrix} X_1 & Y_1 & R_1 \\ X_2 & Y_2 & R_2 \\ \vdots & \vdots & \vdots \\ X_n & Y_n & R_n \end{bmatrix}, \tag{3.1}$$

where every X_j is observed and R_j denotes the response indicator for Y_j, i.e.

$$R_j = \begin{cases} 1, & \text{if } Y_j \text{ is observed,} \\ 0, & \text{otherwise.} \end{cases}$$

If the data are MCAR, then, by definition, it holds that R is independent of (X, Y), and, as a direct consequence, independent of X. Therefore, X and R are uncorrelated, i.e.

$$\mathbb{C}\text{ov}(X, R) = \mathbb{E}(XR) - \mathbb{E}(X)\mathbb{E}(R) = 0.$$

Naturally, for the distribution of (X, Y, R) we can define the parameter $\theta = \mathbb{C}\text{ov}(X, R)$; if $\theta \neq 0$, then X and R are dependent, so the data are not MCAR. If $\theta = 0$, then we can not say that the independence holds, but only that X and R are uncorrelated. However, there are many well-known tests that reject the null hypothesis in such manner (e.g., Kendall's test of independence, Kendall, 1975). This being said, it is natural to construct a test for MCAR based on an estimator of θ. One such estimator is

$$\tilde{T}_n = \frac{1}{n}\sum_{i=1}^{n} X_i R_i - \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)\left(\frac{1}{n}\sum_{i=1}^{n} R_i\right) = \frac{n-1}{n^2}\sum_{i=1}^{n} X_i R_i - \frac{1}{n^2}\sum_{i=1}^{n}\sum_{\substack{j=1 \\ j\neq i}}^{n} X_i R_j. \tag{3.2}$$

$\tilde{T}_n$ is a biased estimator. Indeed, one can easily verify that $\mathbb{E}(\tilde{T}_n) = \frac{n-1}{n}\theta$. After appropriate rescaling, we obtain an unbiased estimator of θ given by:

$$T_n = \frac{1}{n}\sum_{i=1}^{n} X_i R_i - \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{\substack{j=1 \\ j\neq i}}^{n} X_i R_j. \tag{3.3}$$

After some convenient transformations, we obtain:

$$T_n = \frac{1}{\binom{n}{1}}\sum_{i=1}^{n} X_i R_i - \frac{1}{\binom{n}{2}}\sum_{1\leqslant i<j\leqslant n}\sum\frac{1}{2}(X_i R_j + X_j R_i).$$

If we denote

$$U_n^{(1)} = \frac{1}{\binom{n}{1}}\sum_{i=1}^{n}\phi((X_i, Y_i, R_i))$$

and

$$U_n^{(2)} = \frac{1}{\binom{n}{2}}\sum_{1\leqslant i<j\leqslant n}\sum\psi\big((X_i, Y_i, R_i), (X_j, Y_j, R_j)\big),$$

where $\phi((X_i, Y_i, R_i)) = X_i R_i$ and $\psi\big((X_i, Y_i, R_i), (X_j, Y_j, R_j)\big) = (X_i R_j + X_j R_i)/2$, we have that

$$T_n = U_n^{(1)} - U_n^{(2)}.$$

Note that $U_n^{(1)}$ and $U_n^{(2)}$ are U-statistics with kernels ϕ and ψ, respectively. The first of our results is presented below.

THEOREM 3.1 [ALEKSIĆ (2024)]. *Under the null hypothesis of MCAR data and T_n being given in* (3.3)*, it holds that, as $n \to \infty$,*

$$\frac{\sqrt{n}\left(U_n^{(1)} - U_n^{(2)}\right)}{S_n^X S_n^R} \xrightarrow{D} \mathcal{N}(0,1),$$

where $(S_n^X)^2 = \frac{1}{n-1}\sum_{j=1}^n (X_j - \bar{X}_n)^2$ is a sample variance, S_n^X its square root, and S_n^R is defined analogously.

PROOF. We can see that

$$\sigma^2_{1,(1)} := \mathbb{C}\mathrm{ov}(\phi(X_1, Y_1, R_1), \phi(X_1, Y_1, R_1)) = \mathbb{V}\mathrm{ar}\left(\phi(X_1, Y_1, R_1)\right) = \mathbb{E}(X^2R^2) - (\mathbb{E}(X))^2(\mathbb{E}(R))^2.$$

Under the null hypothesis, we have that X and R are independent random variables, and after we note that $R^2 = R$, we obtain

$$\sigma^2_{1,(1)} = \mathbb{E}(X^2)\mathbb{E}(R) - (\mathbb{E}(X))^2(\mathbb{E}(R))^2.$$

Since we are interested in the non-trivial case (not all data observed/missing), we can safely assume that $\mathbb{E}(R) \in (0,1)$, so $\mathbb{E}(R) > (\mathbb{E}(R))^2$. Furthermore, $\mathbb{E}(X^2) > (\mathbb{E}(X))^2$, since only non-degenerate distributions are of interest. These two conditions allow us to conclude that $\sigma^2_{1,(1)} > 0$, i.e. $U_n^{(1)}$ is a non-degenerate U-statistic.

Now we proceed in a similar manner with the statistic $U_n^{(2)}$ and, after similar calculations as before, we obtain:

$$\begin{aligned}\sigma^2_{1,(2)} &:= \mathbb{C}\mathrm{ov}\left(\psi((X_1, Y_1, R_1), (X_2, Y_2, R_2)), \psi((X_1, Y_1, R_1), (X_3, Y_3, R_3))\right) \\ &= \frac{1}{4}\left(\mathbb{E}(R)^2\mathbb{V}\mathrm{ar}(X) + \mathbb{E}(X)^2\mathbb{V}\mathrm{ar}(R)\right).\end{aligned}$$

Again, being interested only in non-trivial cases, we have that $\sigma^2_{1,(2)} > 0$, so $U_n^{(2)}$ is also a non-degenerate U-statistic.

One can readily see that $U_n^{(1)}$ is a U-statistic with expected value $\mathbb{E}\left(\phi(X_1, Y_1, R_1)\right) = \mathbb{E}(XR)$, and that $U_n^{(2)}$ is a U-statistic with expected value $\mathbb{E}\left(\psi((X_1, Y_1, R_1), (X_2, Y_2, R_2))\right) = \mathbb{E}(X)\mathbb{E}(R)$. By Theorem 1.2, we have that

$$\lim_{n\to\infty} n\mathbb{C}\mathrm{ov}\left(U_n^{(1)}, U_n^{(2)}\right) = 2\sigma_{11},$$

where

$$\sigma_{11} = \mathbb{E}\left(\phi_1(X_1, Y_1, R_1), \psi_1(X_1, Y_1, R_1)\right) = \mathbb{C}\mathrm{ov}\left(\phi(X_1, Y_1, R_1), \psi((X_1, Y_1, R_1), (X_2, Y_2, R_2))\right)$$

Under the null hypothesis of MCAR data, we have that:

$$\begin{aligned}\sigma_{11} &= \mathbb{C}\mathrm{ov}\left(X_1R_1, \frac{1}{2}(X_1R_2 + X_2R_1)\right) \\ &= \frac{1}{2}\left(\mathbb{E}\left(X_1R_1(X_1R_2 + X_2R_1)\right) - \mathbb{E}(X_1R_1)\mathbb{E}(X_1R_2 + X_2R_1)\right) \\ &= \frac{1}{2}\left((\mathbb{E}(R))^2\mathbb{V}\mathrm{ar}(X) + (\mathbb{E}(X))^2\mathbb{V}\mathrm{ar}(R)\right).\end{aligned}$$

By Theorem 1.2, we have that as $n \to \infty$, under the null hypothesis,

$$\left(\sqrt{n}\left(U_n^{(1)} - \mathbb{E}(XR)\right), \sqrt{n}\left(U_n^{(2)} - \mathbb{E}(X)\mathbb{E}(R)\right)\right) \tag{3.4}$$

converges in distribution to bivariate normal distribution with zero mean vector and covariance matrix

$$\begin{bmatrix} \mathbb{E}(X^2)\mathbb{E}(R)-(\mathbb{E}(X))^2(\mathbb{E}(R))^2 & (\mathbb{E}(R))^2\mathbb{V}\mathrm{ar}(X)+(\mathbb{E}(X))^2\mathbb{V}\mathrm{ar}(R) \\ (\mathbb{E}(R))^2\mathbb{V}\mathrm{ar}(X)+(\mathbb{E}(X))^2\mathbb{V}\mathrm{ar}(R) & (\mathbb{E}(R))^2\mathbb{V}\mathrm{ar}(X)+(\mathbb{E}(X))^2\mathbb{V}\mathrm{ar}(R) \end{bmatrix}.$$

Applying Continuous Mapping Theorem, we have that the difference $\sqrt{n}(U_n^{(1)}-U_n^{(2)})$ (expectations cancel out under null hypothesis) tends to the difference of the components of the two-dimensional normal distribution of (3.4), i.e.

$$\sqrt{n}\big(U_n^{(1)}-U_n^{(2)}\big) \xrightarrow{D} \mathcal{N}\Big(0, \mathbb{E}(X^2)\mathbb{E}(R)-(\mathbb{E}(X))^2(\mathbb{E}(R))^2 \\ -2\big((\mathbb{E}(R))^2\mathbb{V}\mathrm{ar}(X)+(\mathbb{E}(X))^2\mathbb{V}\mathrm{ar}(R)\big)+(\mathbb{E}(R))^2\mathbb{V}\mathrm{ar}(X)+(\mathbb{E}(X))^2\mathbb{V}\mathrm{ar}(R)\Big),$$

which simplifies to

$$\sqrt{n}\left(U_n^{(1)}-U_n^{(2)}\right) \xrightarrow{D} \mathcal{N}(0, \mathbb{V}\mathrm{ar}(X)\mathbb{V}\mathrm{ar}(R)).$$

In other words,

$$\frac{\sqrt{n}\left(U_n^{(1)}-U_n^{(2)}\right)}{\sqrt{\mathbb{V}\mathrm{ar}(X)\mathbb{V}\mathrm{ar}(R)}} \xrightarrow{D} \mathcal{N}(0,1),$$

as $n \to \infty$. Since the sample standard deviations are consistent estimators, applying Slutsky's theorem, we have that

$$\frac{\sqrt{n}\left(U_n^{(1)}-U_n^{(2)}\right)}{S_n^X S_n^R} \xrightarrow{D} \mathcal{N}(0,1).$$

This concludes the proof. ■

Based on this result, we suggest constructing a MCAR test using test statistic

$$D_n = \frac{\sqrt{n}\left(U_n^{(1)}-U_n^{(2)}\right)}{S_n^X S_n^R},$$

with rejection region given by

$$\left\{|D_n| \geqslant \Phi^{-1}\left(1-\frac{\alpha}{2}\right)\right\} \tag{3.5}$$

at significance level α.

3.1.2 Multivariate data with univariate nonresponse

We next consider the broader situation of p-dimensional observations, where nonresponse occurs in only one variable. Let us suppose that we have the $(p+1)$-variate data, that we observe as the expanded sample

$$\begin{bmatrix} X_1^{(1)} & X_1^{(2)} & \cdots & X_1^{(p)} & Y_1 & R_1 \\ X_2^{(1)} & X_2^{(2)} & \cdots & X_2^{(p)} & Y_2 & R_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ X_n^{(1)} & X_n^{(2)} & \cdots & X_n^{(p)} & Y_n & R_n \end{bmatrix}. \tag{3.6}$$

The natural thing to do in this case is to compute the statistic T_n from (3.3) for every pair $(X^{(j)}, R)$, $j = 1, 2, \ldots, p$. Values near zero suggest that there is little evidence contradicting the MCAR assumption. We now state our next result.

THEOREM 3.2 [ALEKSIĆ (2024)]**.** *Let us have the data represented by the expanded sample* (3.6)*, and let*

$$T_n^{(u)} = \frac{1}{n}\sum_{i=1}^{n} X_i^{(u)} R_i - \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n} X_i^{(u)} R_j,$$

$u = 1,2,\ldots,p$*. Then, under the null hypothesis of MCAR, it holds that*

$$\left(\sqrt{n}\,T_n^{(1)}, \sqrt{n}\,T_n^{(2)}, \ldots, \sqrt{n}\,T_n^{(p)}\right) \xrightarrow{D} \mathcal{N}(0,\Sigma),$$

as $n \to \infty$*, where*

$$\Sigma = \begin{bmatrix} \mathbb{V}\mathrm{ar}\left(X^{(1)}\right)\mathbb{V}\mathrm{ar}(R) & \mathbb{C}\mathrm{ov}\left(X^{(1)}, X^{(2)}\right)\mathbb{V}\mathrm{ar}(R) & \cdots & \mathbb{C}\mathrm{ov}\left(X^{(1)}, X^{(p)}\right)\mathbb{V}\mathrm{ar}(R) \\ \mathbb{C}\mathrm{ov}\left(X^{(1)}, X^{(2)}\right)\mathbb{V}\mathrm{ar}(R) & \mathbb{V}\mathrm{ar}\left(X^{(2)}\right)\mathbb{V}\mathrm{ar}(R) & \cdots & \mathbb{C}\mathrm{ov}\left(X^{(p)}, X^{(2)}\right)\mathbb{V}\mathrm{ar}(R) \\ \vdots & \vdots & \ddots & \vdots \\ \mathbb{C}\mathrm{ov}\left(X^{(1)}, X^{(p)}\right)\mathbb{V}\mathrm{ar}(R) & \mathbb{C}\mathrm{ov}\left(X^{(2)}, X^{(p)}\right)\mathbb{V}\mathrm{ar}(R)\cdots & \cdots & \mathbb{V}\mathrm{ar}\left(X^{(p)}\right)\mathbb{V}\mathrm{ar}(R) \end{bmatrix},$$

i.e.

$$\Sigma = \mathbb{C}\mathrm{ov}\left(\left(X^{(1)}, X^{(2)}, \ldots, X^{(p)}\right)\right)\mathbb{V}\mathrm{ar}(R),$$

where the first term denotes the covariance matrix of a random vector.

PROOF. To keep the expressions simple, we will consider $p = 2$, but it will be obvious that the generalization to the arbitrary p is straightforward.

Assume MCAR and consider that the data are modeled by the random vector $(X^{(1)}, X^{(2)}, Y)$ and, as before, consider the expanded sample

$$\begin{bmatrix} X_1^{(1)} & X_1^{(2)} & Y_1 & R_1 \\ X_2^{(1)} & X_2^{(2)} & Y_2 & R_2 \\ \vdots & \vdots & \vdots & \vdots \\ X_n^{(1)} & X_n^{(2)} & Y_n & R_n \end{bmatrix}. \tag{3.7}$$

Denote

$$T_n^{(1)} = \frac{1}{n}\sum_{i=1}^{n} X_i^{(1)} R_i - \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n} X_i^{(1)} R_j \tag{3.8}$$

and

$$T_n^{(2)} = \frac{1}{n}\sum_{i=1}^{n} X_i^{(2)} R_i - \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n} X_i^{(2)} R_j. \tag{3.9}$$

Certainly, it would be wrong, in the general case, to assume joint normality from the normality of the components. However, since the statistics $\sqrt{n}\,T_n^{(1)}$ and $\sqrt{n}\,T_n^{(2)}$ are (scaled) differences of U-statistics, their joint distribution can be expressed as a linear combination of joint distributions of U-statistics, and hence is asymptotically normal. Given that the asymptotic distribution of $\left(\sqrt{n}\,T_n^{(1)}, \sqrt{n}\,T_n^{(2)}\right)$ is multivariate normal, it will suffice to calculate the limit value of the covariance

$$\mathbb{C}\mathrm{ov}\left(\sqrt{n}\,T_n^{(1)}, \sqrt{n}\,T_n^{(2)}\right) = n\mathbb{C}\mathrm{ov}\left(T_n^{(1)}, T_n^{(2)}\right) = n\mathbb{E}\left(T_n^{(1)} T_n^{(2)}\right),$$

as $n \to \infty$.

Multiplying the expressions (3.8) and (3.9), we obtain

$$\begin{aligned} T_n^{(1)} T_n^{(2)} &= \frac{1}{n^2(n-1)^2} \sum_{i=1}^{n} \sum_{\substack{j=1 \\ j\neq i}}^{n} \sum_{k=1}^{n} \sum_{\substack{l=1 \\ l\neq k}}^{n} X_i^{(1)} R_j X_k^{(2)} R_l + \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} X_i^{(1)} R_i X_j^{(2)} R_j \\ &\quad - \frac{1}{n^2(n-1)} \sum_{i=1}^{n} \sum_{\substack{j=1 \\ j\neq i}}^{n} \sum_{k=1}^{n} X_i^{(1)} R_j X_k^{(2)} R_k - \frac{1}{n^2(n-1)} \sum_{k=1}^{n} \sum_{\substack{l=1 \\ l\neq k}}^{n} \sum_{i=1}^{n} X_k^{(2)} R_l X_i^{(1)} R_i \\ &=: M + N - Q_1 - Q_2. \end{aligned}$$

Having

$$\begin{aligned} n^2(n-1)^2 M &= \sum_i \sum_{j\neq i} \sum_{\substack{l\neq i \\ l\neq j}} X_i^{(1)} R_j X_i^{(2)} R_l + \sum_i \sum_{j\neq i} X_i^{(1)} R_j X_i^{(2)} \\ &\quad + \sum_i \sum_{j\neq i} \sum_{\substack{k\neq i \\ k\neq j}} \sum_{\substack{l\neq i \\ l\neq j \\ l\neq k}} X_i^{(1)} R_j X_k^{(2)} R_l + \sum_i \sum_{j\neq i} \sum_{\substack{k\neq i \\ k\neq j}} X_i^{(1)} R_j X_k^{(2)} R_i + \sum_i \sum_{j\neq i} \sum_{\substack{k\neq i \\ k\neq j}} X_i^{(1)} R_j X_k^{(2)} \\ &\quad + \sum_i \sum_{j\neq i} \sum_{\substack{l\neq i \\ l\neq j}} X_i^{(1)} R_j X_j^{(2)} R_l + \sum_i \sum_{j\neq i} X_i^{(1)} R_j X_j^{(2)} R_i, \end{aligned}$$

we can obtain

$$\begin{aligned} \mathbb{E}\left(n^2(n-1)^2 M\right) &= n(n-1)(n-2)\mathbb{E}\left(X^{(1)}X^{(2)}\right)(\mathbb{E}(R))^2 \\ &\quad + n(n-1)\mathbb{E}\left(X^{(1)}X^{(2)}\right)\mathbb{E}(R) \\ &\quad + n(n-1)(n-2)(n-3)\mathbb{E}\left(X^{(1)}\right)\mathbb{E}\left(X^{(2)}\right)(\mathbb{E}(R))^2 \\ &\quad + n(n-1)(n-2)\mathbb{E}\left(X^{(1)}\right)\mathbb{E}\left(X^{(2)}\right)(\mathbb{E}(R))^2 \\ &\quad + n(n-1)(n-2)\mathbb{E}\left(X^{(1)}\right)\mathbb{E}\left(X^{(2)}\right)\mathbb{E}(R) \\ &\quad + n(n-1)(n-2)\mathbb{E}\left(X^{(1)}\right)\mathbb{E}\left(X^{(2)}\right)(\mathbb{E}(R))^2 \\ &\quad + n(n-1)\mathbb{E}\left(X^{(1)}\right)\mathbb{E}\left(X^{(2)}\right)(\mathbb{E}(R))^2. \end{aligned}$$

Based on this result, we have

$$\begin{aligned} \mathbb{E}(nM) &= \frac{n-2}{n-1}\mathbb{E}\left(X^{(1)}X^{(2)}\right)(\mathbb{E}(R))^2 + \frac{(n-2)(n-3)}{n-1}\mathbb{E}\left(X^{(1)}\right)\mathbb{E}\left(X^{(2)}\right)(\mathbb{E}(R))^2 \\ &\quad + 2\frac{n-2}{n-1}\mathbb{E}\left(X^{(1)}\right)\mathbb{E}\left(X^{(2)}\right)(\mathbb{E}(R))^2 + \frac{n-2}{n-1}\mathbb{E}\left(X^{(1)}\right)\mathbb{E}\left(X^{(2)}\right)(\mathbb{E}(R)) + o(1), \end{aligned}$$

as $n \to \infty$. In a similar manner, it follows that

$$\mathbb{E}(nN) = (n-1)\mathbb{E}\left(X^{(1)}\right)\mathbb{E}\left(X^{(2)}\right)(\mathbb{E}(R))^2 + \mathbb{E}\left(X^{(1)}X^{(2)}\right)\mathbb{E}(R)$$

and

$$\begin{aligned} \mathbb{E}(nQ_1) = \mathbb{E}(nQ_2) &= 2(n-2)\mathbb{E}\left(X^{(1)}\right)\mathbb{E}\left(X^{(2)}\right)(\mathbb{E}(R))^2 + 2\mathbb{E}\left(X^{(1)}X^{(2)}\right)(\mathbb{E}(R))^2 \\ &\quad + 2\mathbb{E}\left(X^{(1)}\right)\mathbb{E}\left(X^{(2)}\right)(\mathbb{E}(R)). \end{aligned}$$

Combining the previous results, and since $\mathbb{E}\left(T_n^{(1)}\right) = \mathbb{E}\left(T_n^{(2)}\right) = 0$ and $R^2 = R$, we have that

$$\lim_{n\to\infty} \mathbb{C}\mathrm{ov}\left(\sqrt{n}T_n^{(1)}, \sqrt{n}T_n^{(2)}\right) = \lim_{n\to\infty} n\mathbb{E}\left(T_n^{(1)}T_n^{(2)}\right)$$

$$
\begin{aligned}
&= \lim_{n\to\infty}(nM + nN - nQ_1 - nQ_2)\\
&= \mathbb{E}\left(X^{(1)}X^{(2)}\right)(\mathbb{E}(R)) - \mathbb{E}\left(X^{(1)}X^{(2)}\right)\mathbb{E}(R)\\
&+ \mathbb{E}\left(X^{(1)}\right)\mathbb{E}\left(X^{(2)}\right)(\mathbb{E}(R))^2 - \mathbb{E}\left(X^{(1)}X^{(2)}\right)(\mathbb{E}(R))^2\\
&= \mathbb{C}\mathrm{ov}\left(X^{(1)}, X^{(2)}\right)\mathbb{E}\left(R^2\right) - \mathbb{C}\mathrm{ov}\left(X^{(1)}, X^{(2)}\right)(\mathbb{E}(R))^2\\
&= \mathbb{C}\mathrm{ov}\left(X^{(1)}, X^{(2)}\right)\mathbb{V}\mathrm{ar}(R).
\end{aligned}
$$

This concludes the proof. ■

Now, let us introduce the following additional assumption: there is no such pair $\left(X^{(u)}, X^{(v)}\right)$ in the vector $\left(X^{(1)}, X^{(2)}, \ldots, X^{(p)}\right)$ such that $\mathbb{C}\mathrm{or}\left(X^{(u)}, X^{(v)}\right) = \pm 1$, for $u, v = 1, 2, \ldots, p$, $u \neq v$. It is well-known (e.g. Strang, 2016, p. 549) that in that case Σ is a regular matrix, and that there exists a matrix $\Sigma^{-1/2}$ such that $\Sigma \cdot \left(\Sigma^{-1/2}\right)^2 = \left(\Sigma^{-1/2}\right)^2 \cdot \Sigma = I$, where I is the identity matrix. Since matrix multiplication is linear, and hence a continuous transformation, we can use Continuous Mapping Theorem to obtain

$$
\left(\Sigma^{-1/2} \cdot \left(\sqrt{n}\,T_n^{(1)}, \sqrt{n}\,T_n^{(2)}, \ldots, \sqrt{n}\,T_n^{(p)}\right)^T\right)^T \xrightarrow{D} \mathcal{N}(0, I),
$$

as $n \to \infty$.

We further assume that each variable $X^{(u)}$, $u = 1, 2, \ldots, p$, has a finite fourth moment. The entries of the matrix Σ are generally unknown and need to be estimated. That can be done using standard bias-adjusted sample covariance matrix, multiplied by bias-adjusted sample variance of R, that are known to be consistent estimators whenever fourth moments of the variables are finite. This gives the estimated matrix $\hat{\Sigma}$. Specifically,

$$
\hat{\Sigma} = \left(\frac{1}{n-1}\sum_{j=1}^{n}(R_j - \bar{R}_j)^2\right)\left(\frac{1}{n-1}\sum_{j=1}^{n}\left(X_j - \bar{X}_n\right)^T\left(X_j - \bar{X}_n\right)\right),
$$

where $X_j = (X_j^{(1)}, \ldots, X_j^{(p)})$, and $\bar{X}_n = \frac{1}{n}\sum_{j=1}^{n} X_j$, and $\bar{R}_n$ is defined similarly. Since matrix inversion and square root are linear and therefore continuous transformations, by the Continuous Mapping Theorem it holds that $\hat{\Sigma}^{-1/2}$ is a consistent estimator of $\Sigma^{-1/2}$. Applying Slutsky's theorem componentwise, we obtain that

$$
\left(A_n^{(1)}, A_n^{(2)}, \ldots, A_n^{(p)}\right) := \left(\hat{\Sigma}^{-1/2} \cdot \left(\sqrt{n}\,T_n^{(1)}, \sqrt{n}\,T_n^{(2)}, \ldots, \sqrt{n}\,T_n^{(p)}\right)^T\right)^T \xrightarrow{D} \mathcal{N}(0, I), \tag{3.10}
$$

as $n \to \infty$.

Values of any $T_n^{(u)}$, $u = 1, 2, \ldots, p$, close to zero indicate that the evidence against MCAR may be weak. The same holds for the components of the vector $\left(A_n^{(1)}, A_n^{(2)}, \ldots, A_n^{(p)}\right)$, since it is just a linear transformation of the previous. Having this conclusion, we construct a test based on the statistic

$$
A_n = \left(A_n^{(1)}\right)^2 + \left(A_n^{(2)}\right)^2 + \cdots + \left(A_n^{(p)}\right)^2,
$$

whose small values indicate that there is not enough evidence agaist the null hypothesis. Having (3.10), we see that A_n is asymptotically distributed as a sum of squares of p IID standard normal variables, and hence A_n is asymptotically χ^2_p-distributed. Finally, we construct a critical region of the test as

$$
\left\{A_n > \chi^2_{p;\alpha}\right\},
$$

where $\chi^2_{p;\alpha}$ is the adequate upper quantile of the χ^2_p distribution.

REMARK 3.1. Note that the statistic A_n, being equal to sum of the squares of the vector

$$\hat{\Sigma}^{-1/2} \cdot \left(\sqrt{n}\, T_n^{(1)}, \ldots, \sqrt{n}\, T_n^{(p)}\right)^T,$$

can be written as

$$A_n = n\left(T_n^{(1)}, T_n^{(2)}, \ldots, T_n^{(p)}\right) \hat{\Sigma}^{-1} \left(T_n^{(1)}, T_n^{(2)}, \ldots, T_n^{(p)}\right)^T.$$

If we, instead, used the MLE $\tilde{\Sigma}$ of the matrix Σ, which is biased, we would have relation $\hat{\Sigma} = \left(\frac{n}{n-1}\right)^2 \tilde{\Sigma}$, and we could write

$$A_n = n\left(\tilde{T}_n^{(1)}, \tilde{T}_n^{(2)}, \ldots, \tilde{T}_n^{(p)}\right) \tilde{\Sigma}^{-1} \left(\tilde{T}_n^{(1)}, \tilde{T}_n^{(2)}, \ldots, \tilde{T}_n^{(p)}\right)^T, \tag{3.11}$$

where $\tilde{T}_n^{(u)}$, $u = 1, 2, \ldots, p$, are defined analogously as in (3.2). Asymptotically, any combination of the previous is equivalent.

3.1.3 General case

As the final step of this section, we soften the assumption of univariate nonresponse, and allow multiple variables to be susceptible to missingness. The only restriction we impose is that there is at least one completely observed variable. Consider the data that can be modeled by random vector $(X^{(1)}, \ldots, X^{(p)}, Y^{(1)}, \ldots, Y^{(q)})$ and the expanded sample

$$\begin{bmatrix} X_1^{(1)} & \cdots & X_1^{(p)} & Y_1^{(1)} & \cdots & Y_1^{(q)} & R_1^{(1)} & \cdots & R_1^{(q)} \\ X_2^{(1)} & \cdots & X_2^{(p)} & Y_2^{(1)} & \cdots & Y_2^{(q)} & R_2^{(1)} & \cdots & R_2^{(q)} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \\ X_n^{(1)} & \cdots & X_n^{(p)} & Y_n^{(1)} & \cdots & Y_n^{(q)} & R_n^{(1)} & \cdots & R_n^{(q)} \end{bmatrix}. \tag{3.12}$$

Suppose that variables $X^{(1)}, \ldots, X^{(p)}$ are completely observed and that $Y^{(1)}, \ldots, Y^{(q)}$ are susceptible to missingness. Similarly to previous subsections, we introduce the statistics

$$T_n^{(u,v)} = \frac{1}{n} \sum_{i=1}^{n} X_i^{(u)} R_i^{(v)} - \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{\substack{j=1 \\ j \neq i}}^{n} X_i^{(u)} R_j^{(v)}, \tag{3.13}$$

$u = 1, 2, \ldots, p$, $v = 1, 2, \ldots, q$. In the exact same manner as Theorem 3.2, we obtain the following theorem. At this point, we omit the proof. First, it is a straightforward generalization of Theorem 3.2. Furthermore, a more general and technically involved result will be presented in Section 3.3, of which this proof will be a special case.

THEOREM 3.3 [ALEKSIĆ (2024)]**.** *Let us have data represented by the expanded sample* (3.12) *and let* $T_n^{(u,v)}$ *be defined as in* (3.13)*, for* $u = 1, 2, \ldots, p$ *and* $v = 1, 2, \ldots, q$*. Then, under the null hypothesis of MCAR, it holds that*

$$\left(\sqrt{n}\, T_n^{(1,1)}, \ldots, \sqrt{n}\, T_n^{(1,q)}, \sqrt{n}\, T_n^{(2,1)}, \ldots, \sqrt{n}\, T_n^{(2,q)}, \ldots, \sqrt{n}\, T_n^{(p,1)}, \ldots, \sqrt{n}\, T_n^{(p,q)}\right) \xrightarrow{D} \mathcal{N}(0, \Sigma), \tag{3.14}$$

as $n \to \infty$*, with*

$$\Sigma = \left[\mathbb{C}\mathrm{ov}\left(X^{(\lceil i/q \rceil)}, X^{(\lceil j/q \rceil)}\right) \mathbb{C}\mathrm{ov}\left(R^{(i \,(\mathrm{mod}\, q))}, R^{(j \,(\mathrm{mod}\, q))}\right)\right]_{i,j \in \{1, \ldots, pq\}}, \tag{3.15}$$

where $a \,(\mathrm{mod}\, b)$ *denotes the remainder of the division of* a *by* b*, and* $\lceil \cdot \rceil$ *is the ceiling function.* ■

REMARK 3.2. In Theorem 3.3, we define $q \,(\mathrm{mod}\, q)$ to be equal to q, and not zero, to get $R^{(q)}$ instead of the non-existent $R^{(0)}$.

As before, if we assume that complete variables have finite fourth moments, $\Sigma^{-1/2}$ exists and can be consistently estimated in a standard way to obtain $\hat{\Sigma}^{-1/2}$ or $\tilde{\Sigma}^{-1/2}$ as a consistent estimator. Finally, it holds that

$$\left(A_n^{(1,1)},\dots,A_n^{(1,q)},A_n^{(2,1)},\dots,A_n^{(2,q)},\dots,A_n^{(p,1)},\dots,A_n^{(p,q)}\right)=\left(\hat{\Sigma}^{-1/2}\cdot\left(\sqrt{n}\,T_n^{(1,1)},\dots,\sqrt{n}\,T_n^{(p,q)}\right)^T\right)^T\xrightarrow{D}\mathcal{N}(0,I),$$

as $n\to\infty$, so we have the convergence

$$A_n=\sum_{u=1}^{p}\sum_{v=1}^{q}\left(A_n^{(u,v)}\right)^2=n\left(T_n^{(1,1)},\dots,T_n^{(p,q)}\right)\hat{\Sigma}^{-1}\left(T_n^{(1,1)},\dots,T_n^{(p,q)}\right)^T\xrightarrow{D}\chi^2_{pq},\tag{3.16}$$

which can be used to calculate the critical values of the test.

Remark 3.3. Note that as long we have p complete and q incomplete variables, the data can be rearranged to take the form (3.12).

Example 3.1. The concise expression (3.15) for Σ can be somewhat abstract, so it could be of a help to discuss it a little bit more. For that purpose, let us have four-dimensional data that consists of the n IID replications of a vector $(X^{(1)},X^{(2)},Y^{(1)},Y^{(2)})$, where we, as usual, consider the variables $X^{(1)}$ and $X^{(2)}$ to be completely observed, and $Y^{(1)}$ and $Y^{(2)}$ partially observed. In that case, Σ has the form

$$\begin{aligned}\Sigma&=\begin{bmatrix}\mathbb{C}\mathrm{ov}(X^{(1)},X^{(1)})\mathbb{C}\mathrm{ov}(R^{(1)},R^{(1)}) & \mathbb{C}\mathrm{ov}(X^{(1)},X^{(1)})\mathbb{C}\mathrm{ov}(R^{(1)},R^{(2)}) & \mathbb{C}\mathrm{ov}(X^{(1)},X^{(2)})\mathbb{C}\mathrm{ov}(R^{(1)},R^{(1)}) & \mathbb{C}\mathrm{ov}(X^{(1)},X^{(2)})\mathbb{C}\mathrm{ov}(R^{(1)},R^{(2)})\\ \mathbb{C}\mathrm{ov}(X^{(1)},X^{(1)})\mathbb{C}\mathrm{ov}(R^{(1)},R^{(2)}) & \mathbb{C}\mathrm{ov}(X^{(1)},X^{(1)})\mathbb{C}\mathrm{ov}(R^{(2)},R^{(2)}) & \mathbb{C}\mathrm{ov}(X^{(1)},X^{(2)})\mathbb{C}\mathrm{ov}(R^{(1)},R^{(2)}) & \mathbb{C}\mathrm{ov}(X^{(1)},X^{(2)})\mathbb{C}\mathrm{ov}(R^{(2)},R^{(2)})\\ \mathbb{C}\mathrm{ov}(X^{(1)},X^{(2)})\mathbb{C}\mathrm{ov}(R^{(1)},R^{(1)}) & \mathbb{C}\mathrm{ov}(X^{(1)},X^{(2)})\mathbb{C}\mathrm{ov}(R^{(1)},R^{(2)}) & \mathbb{C}\mathrm{ov}(X^{(2)},X^{(2)})\mathbb{C}\mathrm{ov}(R^{(1)},R^{(1)}) & \mathbb{C}\mathrm{ov}(X^{(2)},X^{(2)})\mathbb{C}\mathrm{ov}(R^{(1)},R^{(2)})\\ \mathbb{C}\mathrm{ov}(X^{(1)},X^{(2)})\mathbb{C}\mathrm{ov}(R^{(1)},R^{(2)}) & \mathbb{C}\mathrm{ov}(X^{(1)},X^{(2)})\mathbb{C}\mathrm{ov}(R^{(2)},R^{(2)}) & \mathbb{C}\mathrm{ov}(X^{(2)},X^{(2)})\mathbb{C}\mathrm{ov}(R^{(1)},R^{(2)}) & \mathbb{C}\mathrm{ov}(X^{(2)},X^{(2)})\mathbb{C}\mathrm{ov}(R^{(2)},R^{(2)})\end{bmatrix}\\ &=\begin{bmatrix}\mathbb{C}\mathrm{ov}(X^{(1)},X^{(1)})\begin{bmatrix}\mathbb{C}\mathrm{ov}(R^{(1)},R^{(1)}) & \mathbb{C}\mathrm{ov}(R^{(1)},R^{(2)})\\ \mathbb{C}\mathrm{ov}(R^{(1)},R^{(2)}) & \mathbb{C}\mathrm{ov}(R^{(2)},R^{(2)})\end{bmatrix} & \mathbb{C}\mathrm{ov}(X^{(1)},X^{(2)})\begin{bmatrix}\mathbb{C}\mathrm{ov}(R^{(1)},R^{(1)}) & \mathbb{C}\mathrm{ov}(R^{(1)},R^{(2)})\\ \mathbb{C}\mathrm{ov}(R^{(1)},R^{(2)}) & \mathbb{C}\mathrm{ov}(R^{(2)},R^{(2)})\end{bmatrix}\\ \mathbb{C}\mathrm{ov}(X^{(1)},X^{(2)})\begin{bmatrix}\mathbb{C}\mathrm{ov}(R^{(1)},R^{(1)}) & \mathbb{C}\mathrm{ov}(R^{(1)},R^{(2)})\\ \mathbb{C}\mathrm{ov}(R^{(1)},R^{(2)}) & \mathbb{C}\mathrm{ov}(R^{(2)},R^{(2)})\end{bmatrix} & \mathbb{C}\mathrm{ov}(X^{(2)},X^{(2)})\begin{bmatrix}\mathbb{C}\mathrm{ov}(R^{(1)},R^{(1)}) & \mathbb{C}\mathrm{ov}(R^{(1)},R^{(2)})\\ \mathbb{C}\mathrm{ov}(R^{(1)},R^{(2)}) & \mathbb{C}\mathrm{ov}(R^{(2)},R^{(2)})\end{bmatrix}\end{bmatrix}\\ &=\mathbb{C}\mathrm{ov}\left(\left(X^{(1)},X^{(2)}\right)\right)\otimes\mathbb{C}\mathrm{ov}\left(\left(R^{(1)},R^{(2)}\right)\right),\end{aligned}$$

where $\otimes$ denotes the Kronecker product of (covariance) matrices.

Example 3.1 allows us to make straightforward generalization and conclude that the matrix Σ from Theorem 3.3 is equal to the Kronecker product

$$\mathbb{C}\mathrm{ov}\left(\left(X^{(1)},X^{(2)},\dots,X^{(p)}\right)\right)\otimes\mathbb{C}\mathrm{ov}\left(\left(R^{(1)},R^{(2)},\dots,R^{(q)}\right)\right).$$

Since the Kronecker product of two matrices is invertible if and only if both matrices are invertible (e.g., Meyer, 2023, Theorem 2.7.2), we must introduce an additional assumption about the data in order for the test to be well-defined: there should be no perfect linear relationship among the response indicators, as such dependence would imply that their covariance matrix is singular.

To conclude this section, we recapitulate the assumptions required for the test: there is no perfect multicollinearity among the data variables or among the response indicators, and all variables have finite fourth moments.

3.2 A note on the special case of univariate nonresponse

One of the most well-known tests for testing the MCAR assumption is Little's MCAR test, constructed by Little (1988). This test uses a test statistic d^2 that relies on splitting data into groups by missingness patterns. For a sample of IID random vectors with a univariate nonresponse, as in (3.6), the following theorem states that our and Little's statistics coincide.

THEOREM 3.4 [ALEKSIĆ (2024)]. *For the $(p+1)$-variate data with a univariate nonresponse, that can be represented as expanded sample given in* (3.6)*, it holds that*

$$A_n = d^2,$$

where A_n is as in (3.11)*, and d^2 denotes the Little's statistic.*

PROOF. We first need to adapt the expression for Little's statistic to our notation. For simplicity, we discuss the three-dimensional case (3.7), but it will be obvious that the generalization is straightforward.

Denote the vectors

$$L = \left(\frac{1}{\sum_{i=1}^n R_i} \sum_{i=1}^n X_i^{(1)} R_i - \frac{1}{n} \sum_{i=1}^n X_i^{(1)}, \frac{1}{\sum_{i=1}^n R_i} \sum_{i=1}^n X_i^{(2)} R_i - \frac{1}{n} \sum_{i=1}^n X_i^{(2)} \right),$$

$$L_1 = \left(\frac{1}{\sum_{i=1}^n (1-R_i)} \sum_{i=1}^n X_i^{(1)} (1-R_i) - \frac{1}{n} \sum_{i=1}^n X_i^{(1)}, \frac{1}{\sum_{i=1}^n (1-R_i)} \sum_{i=1}^n X_i^{(2)} (1-R_i) - \frac{1}{n} \sum_{i=1}^n X_i^{(2)}, 0 \right)$$

and

$$L_1' = \left(\frac{1}{\sum_{i=1}^n (1-R_i)} \sum_{i=1}^n X_i^{(1)} (1-R_i) - \frac{1}{n} \sum_{i=1}^n X_i^{(1)}, \frac{1}{\sum_{i=1}^n (1-R_i)} \sum_{i=1}^n X_i^{(2)} (1-R_i) - \frac{1}{n} \sum_{i=1}^n X_i^{(2)} \right),$$

and let $\tilde{\Sigma}_1$ be a matrix obtained by expanding $\tilde{\Sigma}$ to include estimates of covariance of $X^{(i)}$, $i = 1,2$, and Y, calculated on those rows i that have Y_i observed. So, the matrix $\tilde{\Sigma}_1$ is of the form

$$\tilde{\Sigma}_1 = \begin{bmatrix} \tilde{\Sigma} & \tilde{\Lambda} \\ \tilde{\Lambda}^T & \tilde{\Delta} \end{bmatrix},$$

where $\tilde{\Lambda}$ and $\tilde{\Delta}$ are estimators of the corresponding matrices, and $\tilde{\Sigma}$ is defined in Remark 3.1, here just for $p = 2$. Denote $\tilde{\Sigma}' = \frac{1}{\bar{R}_n(1-\bar{R}_n)} \tilde{\Sigma}$ and $\tilde{\Sigma}_1' = \frac{1}{\bar{R}_n(1-\bar{R}_n)} \tilde{\Sigma}_1$, where $\bar{R}_n = \frac{1}{n} \sum_{i=1}^n R_i$.

In his paper, Little (1988) introduced the test statistic that, in this special case of a univariate nonresponse, has the form

$$d^2 = n\bar{R}_n L (\tilde{\Sigma}')^{-1} L^T + n(1-\bar{R}_n) L_1 (\tilde{\Sigma}_1')^{-1} L_1^T.$$

But, since in this case the vector L_1 has zero as the last component, matrices $\tilde{\Lambda}$ and $\tilde{\Delta}$ have no effect on d^2, so we can make reduction and obtain that

$$d^2 = n\bar{R}_n L (\tilde{\Sigma}')^{-1} L^T + n(1-\bar{R}_n) L_1' (\tilde{\Sigma}')^{-1} L_1'^T,$$

which, after substituting $\tilde{\Sigma}'$ and $\tilde{\Sigma}_1'$, becomes

$$d^2 = n\bar{R}_n^2 (1-\bar{R}_n) L \tilde{\Sigma}^{-1} L^T + n\bar{R}_n (1-\bar{R}_n)^2 L_1' \tilde{\Sigma}^{-1} L_1'^T. \tag{3.17}$$

Denote $\bar{X}_n^{(1)} = \frac{1}{n} \sum_{i=1}^n X_i^{(1)}$ and $\bar{X}_n^{(2)}$ similarly. Also, let $\overline{X_n^{(1)} R_n} = \frac{1}{n} \sum_{i=1}^n X_i^{(1)} R_i$, and let $\overline{X_n^{(2)} R_n}$ be defined in an analogous manner. Now we can see that

$$L = \frac{1}{\bar{R}_n} \left(\overline{X_n^{(1)} R_n}, \overline{X_n^{(2)} R_n} \right) - \left(\bar{X}_n^{(1)}, \bar{X}_n^{(2)} \right)$$

and

$$L_1' = \frac{1}{(1-\bar{R}_n)}\left(\bar{X}_n^{(1)} - \overline{X_n^{(1)}R_n}, \bar{X}_n^{(2)} - \overline{X_n^{(2)}R_n}\right) - \left(\bar{X}_n^{(1)}, \bar{X}_n^{(2)}\right)$$
$$= \frac{\bar{R}_n}{1-\bar{R}_n}\left(\bar{X}_n^{(1)}, \bar{X}_n^{(2)}\right) - \frac{1}{(1-\bar{R}_n)}\left(\overline{X_n^{(1)}R_n}, \overline{X_n^{(2)}R_n}\right).$$

Next, we calculate both of the terms in (3.17). First, we have that

$$\begin{aligned}\bar{R}_n^2(1-\bar{R}_n)L\tilde{\Sigma}^{-1}L^T = n(1-\bar{R}_n)\left(\overline{X_n^{(1)}R_n}, \overline{X_n^{(2)}R_n}\right)\tilde{\Sigma}^{-1}\left(\overline{X_n^{(1)}R_n}, \overline{X_n^{(2)}R_n}\right)^T \\ -2n\bar{R}_n(1-\bar{R}_n)\left(\overline{X_n^{(1)}R_n}, \overline{X_n^{(2)}R_n}\right)\tilde{\Sigma}^{-1}\left(\bar{X}_n^{(1)}, \bar{X}_n^{(2)}\right)^T \\ +n\bar{R}_n^2(1-\bar{R}_n)\left(\bar{X}_n^{(1)}, \bar{X}_n^{(2)}\right)\tilde{\Sigma}^{-1}\left(\bar{X}_n^{(1)}, \bar{X}_n^{(2)}\right)^T.\end{aligned}$$

Then, it holds that

$$\begin{aligned}n\bar{R}_n(1-\bar{R}_n)^2L_1'\tilde{\Sigma}^{-1}L_1'^T = n\bar{R}_n\left(\overline{X_n^{(1)}R_n}, \overline{X_n^{(2)}R_n}\right)\tilde{\Sigma}^{-1}\left(\overline{X_n^{(1)}R_n}, \overline{X_n^{(2)}R_n}\right)^T \\ -2n\bar{R}_n^2\left(\overline{X_n^{(1)}R_n}, \overline{X_n^{(2)}R_n}\right)\tilde{\Sigma}^{-1}\left(\bar{X}_n^{(1)}, \bar{X}_n^{(2)}\right)^T \\ +n\bar{R}_n^3\left(\bar{X}_n^{(1)}, \bar{X}_n^{(2)}\right)\tilde{\Sigma}^{-1}\left(\bar{X}_n^{(1)}, \bar{X}_n^{(2)}\right)^T.\end{aligned}$$

Combining, we have

$$\begin{aligned}d^2 = n\left(\overline{X_n^{(1)}R_n}, \overline{X_n^{(2)}R_n}\right)\tilde{\Sigma}^{-1}\left(\overline{X_n^{(1)}R_n}, \overline{X_n^{(2)}R_n}\right)^T \\ -2n\bar{R}_n\left(\overline{X_n^{(1)}R_n}, \overline{X_n^{(2)}R_n}\right)\tilde{\Sigma}^{-1}\left(\bar{X}_n^{(1)}, \bar{X}_n^{(2)}\right)^T \\ +n\bar{R}_n^2\left(\bar{X}_n^{(1)}, \bar{X}_n^{(2)}\right)\tilde{\Sigma}^{-1}\left(\bar{X}_n^{(1)}, \bar{X}_n^{(2)}\right)^T.\end{aligned}$$

On the other hand, from (3.11) (for $p = 2$) and from the fact that $\tilde{T}_n^{(u)} = \overline{X_n^{(u)}R_n} - \bar{X}_n^{(u)}\bar{R}_n$, $u = 1,2$, it follows that

$$\begin{aligned}A_n &= n\left(\tilde{T}_n^{(1)}, \tilde{T}_n^{(2)}\right)\tilde{\Sigma}^{-1}\left(\tilde{T}_n^{(1)}, \tilde{T}_n^{(2)}\right)^T \\ &= n\left(\overline{X_n^{(1)}R_n}, \overline{X_n^{(2)}R_n}\right)\tilde{\Sigma}^{-1}\left(\overline{X_n^{(1)}R_n}, \overline{X_n^{(2)}R_n}\right)^T \\ &\quad -2n\bar{R}_n\left(\overline{X_n^{(1)}R_n}, \overline{X_n^{(2)}R_n}\right)\tilde{\Sigma}^{-1}\left(\bar{X}_n^{(1)}, \bar{X}_n^{(2)}\right)^T \\ &\quad +n\bar{R}_n^2\left(\bar{X}_n^{(1)}, \bar{X}_n^{(2)}\right)\tilde{\Sigma}^{-1}\left(\bar{X}_n^{(1)}, \bar{X}_n^{(2)}\right)^T,\end{aligned}$$

which is exactly d^2. Thus, Theorem 3.4 holds. ■

3.3 Generalization

The A_n-based test left some properties to be desired. The first major drawback of the test is that it can be used only on a dataset that has at least one complete column. The second and more significant limitation is that it entirely disregards the partially observed variables $Y^{(1)}, \ldots, Y^{(q)}$. This leads not only to a loss of power but also to a potential inability to detect alternatives to MCAR where response indicators depend on $Y^{(1)}, \ldots, Y^{(q)}$, but not on $X^{(1)}, \ldots, X^{(p)}$. This setting is not very uncommon; e.g., it can occur whenever response indicators depend on incomplete variables, but complete and incomplete variables are mutually independent. Therefore, it is of

essential importance to address this issue. In this section, we improve the test so it becomes able to utilize the partially observed variables.

Once again, we consider the expanded sample (3.12). Not to get confused, let us introduce only a slightly different notation than one from previous section. Statistic $T_n^{(u,v)}$ from (3.13) computed from variable pair $\left(X^{(u)}, R^{(v)}\right)$ will now be denoted $T_{n,X}^{(u,v)}$.

Our main goal is to find a way to compute the statistic $T_n^{(u,v)}$ from (3.13) for each pair $(Y^{(u)}, R^{(v)})$, with $u \neq v$, in order to obtain an estimate of the covariance between them. The subsequent goal is to use those statistics to extend the vector from (3.16) to include them, and, as a consequence, to expand the set of detectable alternatives of the test. Under the null hypothesis of MCAR data, it is reasonable to calculate it on those cases where $Y^{(u)}$ is observed. In that case, the statistic can be written as

$$T_{n,Y}^{(u,v)} = \frac{1}{\hat{n}^{(u)}} \sum_{i=1}^{n} Y_i^{(u)} R_i^{(u)} R_i^{(v)} - \frac{1}{\hat{n}^{(u)}(n-1)} \sum_{i=1}^{n} \sum_{\substack{j=1 \\ j \neq i}}^{n} Y_i^{(u)} R_i^{(u)} R_j^{(v)}, \quad 1 \leqslant u, v \leqslant q, \quad u \neq v, \tag{3.18}$$

where

$$\hat{n}^{(u)} = \sum_{i=1}^{n} R_i^{(u)}, \quad 1 \leqslant u \leqslant q.$$

Remark 3.4. We note that the form (3.18) is a generalization of (3.13), since, for complete variables, all of the response indicators are equal to 1.

It is intuitive (and true) that, under MCAR data, a non-degenerate U-statistic computed from complete cases, appropriately normalized, has the same asymptotic distribution as the one based on the complete sample. However, rigorous proof was anything but trivial, as we will see in the Chapter 4, where we will present it. The complexity of the result is due to the fact that $\hat{n}^{(u)}$ is not constant, but random. Formally speaking, $T_{n,Y}^{(u,v)}$ is not a U-statistic, but is asymptotically equivalent to one. In our case, which involves the difference between two test statistics and the joint distribution of such statistics, we should not expect the situation to be any simpler.

The main reason for introducing the statistic $T_{n,Y}^{(u,v)}$ from (3.18) is that it serves as an unbiased estimator of $\mathbb{C}\mathrm{ov}\left(Y^{(u)}, R^{(v)}\right)$, which is a measure of dependence between $Y^{(u)}$ and $R^{(v)}$. However, the estimate of any value proportional to it would also suffice. So, naturally, one could think of

$$\hat{T}_{n,Y}^{(u,v)} = \frac{1}{n} \sum_{i=1}^{n} Y_i^{(u)} R_i^{(u)} R_i^{(v)} - \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{\substack{j=1 \\ j \neq i}}^{n} Y_i^{(u)} R_i^{(u)} R_j^{(v)}. \tag{3.19}$$

as a modification of the test statistic. This choice seems appropriate, since we use deterministic n instead of random $\hat{n}^{(u)}$, and

$$\hat{T}_{n,Y}^{(u,v)} = \frac{\hat{n}^{(u)}}{n} T_{n,Y}^{(u,v)},$$

so it seems like Slutsky's theorem could be used as a final step of deriving the asymptotic behavior. However, a more fundamental problem lies beneath the surface. Under mutual uncorrelatedness of response indicators, it holds that

$$\mathbb{E}\left(\hat{T}_{n,Y}^{(u,v)}\right) = \mathbb{E}\left(T_{n,Y}^{(u,v)}\right)\mathbb{E}\left(R^{(u)}\right) = \mathbb{C}\mathrm{ov}\left(Y^{(u)}, R^{(v)}\right)\mathbb{E}\left(R^{(u)}\right),$$

which is equal to zero under MCAR. This makes $\hat{T}_{n,Y}^{(u,v)}$ appear to be a suitable choice, as its expected value is proportional to the target covariance. However, if response indicators exhibit any form of dependence among themselves, the statistic $T_{n,Y}^{(u,v)}$ from (3.18) is no longer a

complete-case estimate under MCAR, since the data on which it is computed are not MCAR in that case. Indeed, the data here consist of realizations of the pair $(Y^{(u)}, R^{(v)})$, while complete cases are selected with respect to $R^{(u)}$. These response indicators corresponding to different variables need not be mutually independent for the data to be MCAR; the only assumption required by the test is the absence of multicollinearity among them, that is, their covariance matrix must be regular.

If we assume MCAR and make no additional assumptions about the response indicators, it is straightforward to see that

$$\mathbb{E}\big(\hat{T}_{n,Y}^{(u,v)}\big) = \mathbb{E}\big(Y^{(u)}\big)\mathbb{C}\mathrm{ov}\big(R^{(u)}, R^{(v)}\big),$$

which does not necessarily equal zero under the null hypothesis, as it should be. In fact, $\hat{T}_{n,Y}^{(u,v)}$ is a U-statistic that estimates $\mathbb{C}\mathrm{ov}\big(Y^{(u)}R^{(u)}, R^{(v)}\big)$, which may or may not be close to desired $\mathbb{C}\mathrm{ov}\big(Y^{(u)}, R^{(v)}\big)\,\mathbb{E}\big(R^{(u)}\big)$, depending on the internal structure of the response indicators. In other words, the statistic $\hat{T}_{n,Y}^{(u,v)}$ can be interpreted as an indirect measure of the covariance between the incomplete variable and the response indicator, although with variable reliability.

To be able to examine this issue in more detail, we first need asymptotic results for these statistics.

Lemma 3.1. *Under MCAR, it holds that*

$$\lim_{n\to\infty} \mathbb{C}\mathrm{ov}\big(\sqrt{n}\,\hat{T}_{n,Y}^{(u,v)},\ \sqrt{n}\,\hat{T}_{n,Y}^{(r,s)}\big) = \mathbb{E}\big(Y^{(u)}Y^{(r)}\big)\cdot A + \mathbb{E}\big(Y^{(u)}\big)\mathbb{E}\big(Y^{(r)}\big)\cdot B, \tag{3.20}$$

where

$$\begin{aligned} A = {} & \mathbb{E}\big(R^{(u)}R^{(v)}R^{(r)}R^{(s)}\big) - \mathbb{E}\big(R^{(u)}R^{(v)}R^{(r)}\big)\mathbb{E}\big(R^{(s)}\big) \\ & - \mathbb{E}\big(R^{(u)}R^{(r)}R^{(s)}\big)\mathbb{E}\big(R^{(v)}\big) + \mathbb{E}\big(R^{(u)}R^{(r)}\big)\mathbb{E}\big(R^{(v)}\big)\mathbb{E}\big(R^{(s)}\big) \end{aligned}$$

and

$$\begin{aligned} B = {} & \mathbb{E}\big(R^{(u)}R^{(s)}\big)\mathbb{E}\big(R^{(v)}\big)\mathbb{E}\big(R^{(r)}\big) + \mathbb{E}\big(R^{(v)}R^{(r)}\big)\mathbb{E}\big(R^{(u)}\big)\mathbb{E}\big(R^{(s)}\big) + \mathbb{E}\big(R^{(v)}R^{(s)}\big)\mathbb{E}\big(R^{(u)}\big)\mathbb{E}\big(R^{(r)}\big) \\ & + 2\mathbb{E}\big(R^{(r)}R^{(s)}\big)\mathbb{E}\big(R^{(u)}\big)\mathbb{E}\big(R^{(v)}\big) + 2\mathbb{E}\big(R^{(u)}R^{(v)}\big)\mathbb{E}\big(R^{(r)}\big)\mathbb{E}\big(R^{(s)}\big) - \mathbb{E}\big(R^{(u)}R^{(v)}R^{(s)}\big)\mathbb{E}\big(R^{(r)}\big) \\ & - 4\mathbb{E}\big(R^{(u)}\big)\mathbb{E}\big(R^{(v)}\big)\mathbb{E}\big(R^{(r)}\big)\mathbb{E}\big(R^{(s)}\big) - \mathbb{E}\big(R^{(u)}R^{(v)}\big)\mathbb{E}\big(R^{(r)}R^{(s)}\big) - \mathbb{E}\big(R^{(v)}R^{(r)}R^{(s)}\big)\mathbb{E}\big(R^{(u)}\big). \end{aligned}$$

In particular, if either the incomplete variables have zero means or the response indicators are uncorrelated, it holds that

$$\lim_{n\to\infty} \mathbb{C}\mathrm{ov}\big(\sqrt{n}\,\hat{T}_{n,Y}^{(u,v)},\ \sqrt{n}\,\hat{T}_{n,Y}^{(r,s)}\big) = \mathbb{E}\big(Y^{(u)}Y^{(r)}\big)\cdot A. \tag{3.21}$$

Proof. Let u, v, r, s be fixed. Begin by noting that

$$\hat{T}_{n,Y}^{(u,v)} = \frac{1}{\binom{n}{1}}\sum_{i=1}^{n} Y_i^{(u)}R_i^{(u)}R_i^{(v)} - \frac{1}{\binom{n}{2}}\sum\sum_{1\leqslant i<j\leqslant n}\frac{1}{2}\Big(Y_i^{(u)}R_i^{(u)}R_j^{(v)} + Y_j^{(u)}R_j^{(u)}R_i^{(v)}\Big) =: M_n - N_n$$

and, similarly

$$\hat{T}_{n,Y}^{(r,s)} = \frac{1}{\binom{n}{1}}\sum_{i=1}^{n} Y_i^{(r)}R_i^{(r)}R_i^{(s)} - \frac{1}{\binom{n}{2}}\sum\sum_{1\leqslant i<j\leqslant n}\frac{1}{2}\Big(Y_i^{(r)}R_i^{(r)}R_j^{(s)} + Y_j^{(r)}R_j^{(r)}R_i^{(s)}\Big) =: Q_n - S_n.$$

Relying on Theorem 1.2, we have that

$$\lim_{n\to\infty} n\mathbb{C}\mathrm{ov}(M_n, Q_n) = 1\cdot 1\cdot \mathbb{C}\mathrm{ov}\big(Y_1^{(u)}R_1^{(u)}R_1^{(v)},\ Y_1^{(r)}R_1^{(r)}R_1^{(s)}\big)$$

$$= \mathbb{E}\big(Y^{(u)}Y^{(r)}\big)\mathbb{E}\big(R^{(u)}R^{(v)}R^{(r)}R^{(s)}\big) - \mathbb{E}\big(Y^{(u)}\big)\mathbb{E}\big(Y^{(r)}\big)\mathbb{E}\big(R^{(u)}R^{(v)}\big)\mathbb{E}\big(R^{(r)}R^{(s)}\big),$$

and, similarly,

$$\begin{aligned}
&\lim_{n\to\infty} n\mathbb{C}\mathrm{ov}(M_n, S_n) \\
&\quad = 1\cdot 2\cdot \mathbb{C}\mathrm{ov}\Big(Y_1^{(u)}R_1^{(u)}R_1^{(v)}, \frac{1}{2}\big(Y_1^{(r)}R_1^{(r)}R_2^{(s)} + Y_2^{(r)}R_2^{(r)}R_1^{(s)}\big)\Big) \\
&\quad = \mathbb{C}\mathrm{ov}\big(Y_1^{(u)}R_1^{(u)}R_1^{(v)},\, Y_1^{(r)}R_1^{(r)}R_2^{(s)}\big) + \mathbb{C}\mathrm{ov}\big(Y_1^{(u)}R_1^{(u)}R_1^{(v)},\, Y_2^{(r)}R_2^{(r)}R_1^{(s)}\big) \\
&\quad = \mathbb{E}\big(Y^{(u)}Y^{(r)}\big)\mathbb{E}\big(R^{(u)}R^{(v)}R^{(r)}\big)\mathbb{E}\big(R^{(s)}\big) - \mathbb{E}\big(Y^{(u)}\big)\mathbb{E}\big(Y^{(r)}\big)\mathbb{E}\big(R^{(u)}R^{(v)}\big)\mathbb{E}\big(R^{(r)}\big)\mathbb{E}\big(R^{(s)}\big) \\
&\qquad + \mathbb{E}\big(Y^{(u)}\big)\mathbb{E}\big(Y^{(r)}\big)\mathbb{E}\big(R^{(u)}R^{(v)}R^{(s)}\big)\mathbb{E}\big(R^{(r)}\big) - \mathbb{E}\big(Y^{(u)}\big)\mathbb{E}\big(Y^{(r)}\big)\mathbb{E}\big(R^{(u)}R^{(v)}\big)\mathbb{E}\big(R^{(r)}\big)\mathbb{E}\big(R^{(s)}\big) \\
&\quad = \mathbb{E}\big(Y^{(u)}Y^{(r)}\big)\mathbb{E}\big(R^{(u)}R^{(v)}R^{(r)}\big)\mathbb{E}\big(R^{(s)}\big) - 2\mathbb{E}\big(Y^{(u)}\big)\mathbb{E}\big(Y^{(r)}\big)\mathbb{E}\big(R^{(u)}R^{(v)}\big)\mathbb{E}\big(R^{(r)}\big)\mathbb{E}\big(R^{(s)}\big) \\
&\qquad + \mathbb{E}\big(Y^{(u)}\big)\mathbb{E}\big(Y^{(r)}\big)\mathbb{E}\big(R^{(u)}R^{(v)}R^{(s)}\big)\mathbb{E}\big(R^{(r)}\big).
\end{aligned}$$

Analogously, we obtain

$$\begin{aligned}
&\lim_{n\to\infty} n\mathbb{C}\mathrm{ov}(N_n, Q_n) \\
&\quad = 2\cdot 1\cdot \mathbb{C}\mathrm{ov}\Big(\frac{1}{2}\big(Y_1^{(u)}R_1^{(u)}R_2^{(v)} + Y_2^{(u)}R_2^{(u)}R_1^{(v)}\big),\, Y_1^{(r)}R_1^{(r)}R_1^{(s)}\Big) \\
&\quad = \mathbb{C}\mathrm{ov}\big(Y_1^{(u)}R_1^{(u)}R_2^{(v)}, Y_1^{(r)}R_1^{(r)}R_1^{(s)}\big) + \mathbb{C}\mathrm{ov}\big(Y_2^{(u)}R_2^{(u)}R_1^{(v)}, Y_1^{(r)}R_1^{(r)}R_1^{(s)}\big)\mathbb{E}\big(R^{(v)}\big) \\
&\quad = \mathbb{E}\big(Y^{(u)}Y^{(r)}\big)\mathbb{E}\big(R^{(u)}R^{(r)}R^{(s)}\big) + \mathbb{E}\big(Y^{(u)}\big)\mathbb{E}\big(Y^{(r)}\big)\mathbb{E}\big(R^{(v)}R^{(r)}R^{(s)}\big)\mathbb{E}\big(R^{(u)}\big) \\
&\qquad - 2\mathbb{E}\big(Y^{(u)}\big)\mathbb{E}\big(Y^{(r)}\big)\mathbb{E}\big(R^{(u)}R^{(v)}\big)\mathbb{E}\big(R^{(r)}R^{(s)}\big)
\end{aligned}$$

and

$$\begin{aligned}
&\lim_{n\to\infty} n\mathbb{C}\mathrm{ov}(N_n, S_n) \\
&\quad = 2\cdot 2\cdot \mathbb{C}\mathrm{ov}\Big(\frac{1}{2}\big(Y_1^{(u)}R_1^{(u)}R_2^{(v)} + Y_2^{(u)}R_2^{(u)}R_1^{(v)}\big), \frac{1}{2}\big(Y_1^{(r)}R_1^{(r)}R_3^{(s)} + Y_3^{(r)}R_3^{(r)}R_1^{(s)}\big)\Big) \\
&\quad = \mathbb{C}\mathrm{ov}\big(Y_1^{(u)}R_1^{(u)}R_2^{(v)},\, Y_1^{(r)}R_1^{(r)}R_3^{(s)}\big) + \mathbb{C}\mathrm{ov}\big(Y_1^{(u)}R_1^{(u)}R_2^{(v)},\, Y_3^{(r)}R_3^{(r)}R_1^{(s)}\big) \\
&\qquad + \mathbb{C}\mathrm{ov}\big(Y_2^{(u)}R_2^{(u)}R_1^{(v)},\, Y_1^{(r)}R_1^{(r)}R_3^{(s)}\big) + \mathbb{C}\mathrm{ov}\big(Y_2^{(u)}R_2^{(u)}R_1^{(v)},\, Y_3^{(r)}R_3^{(r)}R_1^{(s)}\big) \\
&\quad = \mathbb{E}\big(Y^{(u)}Y^{(r)}\big)\mathbb{E}\big(R^{(u)}R^{(r)}\big)\mathbb{E}\big(R^{(v)}\big)\mathbb{E}\big(R^{(s)}\big) + \mathbb{E}\big(Y^{(u)}\big)\mathbb{E}\big(Y^{(r)}\big)\mathbb{E}\big(R^{(u)}R^{(s)}\big)\mathbb{E}\big(R^{(v)}\big)\mathbb{E}\big(R^{(r)}\big) \\
&\qquad + \mathbb{E}\big(Y^{(u)}\big)\mathbb{E}\big(Y^{(r)}\big)\mathbb{E}\big(R^{(v)}R^{(r)}\big)\mathbb{E}\big(R^{(u)}\big)\mathbb{E}\big(R^{(s)}\big) + \mathbb{E}\big(Y^{(u)}\big)\mathbb{E}\big(Y^{(r)}\big)\mathbb{E}\big(R^{(v)}R^{(s)}\big)\mathbb{E}\big(R^{(u)}\big)\mathbb{E}\big(R^{(r)}\big) \\
&\qquad - 4\mathbb{E}\big(Y^{(u)}\big)\mathbb{E}\big(Y^{(r)}\big)\mathbb{E}\big(R^{(u)}\big)\mathbb{E}\big(R^{(v)}\big)\mathbb{E}\big(R^{(r)}\big)\mathbb{E}\big(R^{(s)}\big).
\end{aligned}$$

By noting that

$$\begin{aligned}
\lim_{n\to\infty} n\mathbb{C}\mathrm{ov}\big(\hat{T}_{n,Y}^{(u,v)}, \hat{T}_{n,Y}^{(r,s)}\big) &= \lim_{n\to\infty} n\mathbb{C}\mathrm{ov}(M_n, Q_n) - \lim_{n\to\infty} n\mathbb{C}\mathrm{ov}(M_n, S_n) \\
&\quad - \lim_{n\to\infty} n\mathbb{C}\mathrm{ov}(N_n, Q_n) + \lim_{n\to\infty} n\mathbb{C}\mathrm{ov}(N_n, S_n)
\end{aligned}$$

and combining the derived expressions, we obtain the statement of the Lemma. This concludes the proof. ■

COROLLARY 3.4.1. *Under MCAR, it holds that*

$$\begin{aligned}
&\lim_{n\to\infty} \mathbb{C}\mathrm{ov}\big(\sqrt{n}\,T_{n,X}^{(u,v)}, \sqrt{n}\,\hat{T}_{n,Y}^{(r,s)}\big) \\
&\quad = \mathbb{C}\mathrm{ov}\big(X^{(u)}, Y^{(r)}\big)\Big(\mathbb{E}\big(R^{(v)}R^{(r)}R^{(s)}\big) + \mathbb{E}\big(R^{(v)}\big)\mathbb{E}\big(R^{(r)}\big)\mathbb{E}\big(R^{(s)}\big) \\
&\qquad - \mathbb{E}\big(R^{(v)}\big)\mathbb{E}\big(R^{(r)}R^{(s)}\big) - \mathbb{E}\big(R^{(v)}R^{(r)}\big)\mathbb{E}\big(R^{(s)}\big)\Big). \qquad (3.22)
\end{aligned}$$

PROOF. The results follows from Lemma 3.4.1 by setting $Y^{(u)} = X^{(u)}$ and noting that $R^{(u)} \equiv 1$. ■

COROLLARY 3.4.2. *Under MCAR, it holds that*

$$\lim_{n\to\infty} \mathbb{C}\mathrm{ov}\left(\sqrt{n}\,T_{n,X}^{(u,v)}, \sqrt{n}\,T_{n,X}^{(r,s)}\right) = \mathbb{C}\mathrm{ov}\left(X^{(u)}, X^{(r)}\right)\mathbb{C}\mathrm{ov}\left(R^{(v)}, R^{(s)}\right). \tag{3.23}$$

PROOF. Follows directly from Corollary 3.4.1 by setting $Y^{(r)} = X^{(r)}$ and noting that $R^{(r)} \equiv 1$. ■

REMARK 3.5. Note that Corollary 3.4.2 is in fact Theorem 3.3. As stated in Subsection 3.1.3, it is a corollary of a more general result.

The following result summarizes our findings.

THEOREM 3.5 [ALEKSIĆ (2025)]**.** *Assume the data are MCAR, all variables have finite fourth moments, and either the incomplete variables have zero means or the response indicators are uncorrelated. Then*

$$\sqrt{n}\Bigg(T_{n,X}^{(1,1)}, \ldots, T_{n,X}^{(1,q)}, T_{n,X}^{(2,1)}, \ldots, T_{n,X}^{(2,q)}, \ldots, T_{n,X}^{(p,1)}, \ldots, T_{n,X}^{(p,q)}, \\ \hat{T}_{n,Y}^{(1,2)}, \ldots, \hat{T}_{n,Y}^{(1,q)}, \hat{T}_{n,Y}^{(2,1)}, \hat{T}_{n,Y}^{(2,3)}, \ldots, \hat{T}_{n,Y}^{(2,q)}, \ldots, \hat{T}_{n,Y}^{(q,1)}, \ldots, \hat{T}_{n,Y}^{(q,q-1)}\Bigg) \xrightarrow{D} \mathcal{N}(0, \Lambda)$$

and

$$A_n' := n\left(T_{n,X}^{(1,1)}, \ldots, \hat{T}_{n,Y}^{(q,q-1)}\right)\hat{\Lambda}^{-1}\left(T_{n,X}^{(1,1)}, \ldots, \hat{T}_{n,Y}^{(q,q-1)}\right)^T \xrightarrow{D} \chi^2_{pq+q(q-1)}, \tag{3.24}$$

where Λ is corresponding limiting covariance matrix with limiting covariances (3.21)*,* (3.22)*, and* (3.23)*, and $\hat{\Lambda}$ is its standard bias-adjusted estimate.*

PROOF. The results follows directly from the equations (3.21), (3.22), and (3.23), and the fact that the $\hat{\Lambda}$ is a consistent estimator under the finite fourth moments assumption. ■

The convergence (3.24) can be subsequently used to construct the rejection region and compute the p-value of the improved MCAR test.

REMARK 3.6. The assumption of either the incomplete variables having zero means or the response indicators being uncorrelated makes the test inapplicable in most scenarios, so we mitigate this issue by centering the data before conducting the test. Since we center them using (complete-case where necessary) estimates of their means rather than the (unknown) theoretical ones, this introduces a potential methodological concern. One of the goals of the empirical study is to examine the robustness of the test with respect to this issue.

3.4 Empirical study

In this section, we present the results of an extensive simulation study which is conducted to examine how the novel test behaves in terms of empirical type I error and power. The novel test is compared to Little's MCAR test, as well as the old test based on the statistic A_n from (3.16), which novel A_n' from (3.24) improves upon.

As previously noted, the original test based on A_n could only detect correlations between the response indicators and fully observed variables. Therefore, in addition to evaluating alternatives undetectable by the original test, we also compare A_n, Little's d^2, and the proposed A_n' in these scenarios (Aleksić, 2024). This allows us to assess whether the improvements offered by the new test come with potential trade-offs, such as reduced power or calibration issues.

The missingness probability for any value in the data, i.e. the probability that a specific data cell is missing, ranges from 3% to 30%; we have decided to study sample sizes of $n = 100$,

$n = 200$, and $n = 300$, which appears to be adequate for illustrating the quality of asymptotic approximations.

Throughout the rest of this section, we use the abbreviation $iXjY$ to denote the dataset with a total of $i + j$ variables, where i of them are complete, and j of them are incomplete. We present only the $2X3Y$ results in the main text to avoid overload; findings for other cases are consistent and thus omitted.

All simulations are performed with $N = 2000$ replications and at nominal level of $\alpha = 0.05$.

3.4.1 Study design

Generating the data

For the data distributions, we use the standard normal distribution, as well as the normal distribution with marginal means equal to 1 and covariance matrix $0.5I + 0.5J$, where I is the identity matrix and J is a matrix with all elements equal to 1. We also consider a Clayton copula with parameter 1 and exponential $\mathscr{E}(1)$ margins, as well as χ^2_4 margins (see, e.g., Fischer and Köck, 2012). The main idea behind this choice is that Little's test relies on the normality assumption. Given this, it is important to assess the performance of the novel test for normally distributed data, for data whose distribution deviates substantially from normality, and intermediate cases. The Clayton copula was also used in an independence testing scenario by Cuparić and Milošević (2024), where the test of independence by Kochar and Gupta (1990) was adapted for the setting of randomly censored data.

Generating missingness with uncorrelated response indicators

For implementation, `R` package `missMethods` is used (Rockel, 2023). For the null distribution case, function `delete_MCAR` is used. We stick to the alternatives implemented in functions `delete_MAR_1_to_x`, with recommended choice of $x = 9$ and argument `n_mis_stochastic = FALSE`, and `delete_MAR_rank`. The main idea between these mechanisms is that, for each incomplete data column, we have the so-called *control column*, that is fully observed, and the data from that column is used to dictate the missingness probability in the incomplete one. The first mechanism works by setting a specific threshold (default is median), and then splitting the cases into two groups: those that have value of control variable smaller than the threshold, and those that do not. Then, the missingness is introduced such that the odds of a value being missing in those two groups are $1 : x$. For the second mechanism, the probability that a value is missing is directly proportional on the rank of its observed pair. For much more details, we refer to Santos et al. (2019), where they were first introduced.

Generating missingness with correlated response indicators

To examine the behavior of the improved test when centering is required, we generate missing data with correlated response indicators, controlling their correlation coefficient. This is done by modifying the functions `delete_MCAR`, `delete_MAR_1_to_x`, and `delete_MAR_rank`. Algorithm 3.1 illustrates the procedure for the $2X3Y$ data and positive correlation, and it can be easily adapted to other dimensions and correlation structures.

The following lemma formally establishes that the correlated response indicators generated in Algorithm 3.1 have a correlation coefficient equal to ρ.

Lemma 3.2. *Let R_1 and R_2 be two independent indicator random variables with the same expected value q. Let U be a random variable uniformly distributed on $[0,1]$ and independent of both R_1 and R_2. If, for $\rho \in [0,1]$,*

$$R_3 = I\left\{U \leqslant \rho\right\} R_1 + I\left\{U > \rho\right\} R_2,$$

Algorithm 3.1 Generating missingness with positively correlated response indicators.

1: Start with the complete sample $(x_1, x_2, \ldots, x_n)$, where each $x_j = \left(x_j^{(1)}, x_j^{(2)}, y_j^{(1)}, y_j^{(2)}, y_j^{(3)}\right)$;
2: Specify the desired missingness probability p and the correlation coefficient ρ;
3: Generate missingness in variables $y^{(1)}$ and $y^{(2)}$ using the probability p and a chosen method; record the response indicator vector $r^{(2)}$;
4: Generate a random vector r of length n consisting of zeros and ones, where 0 appears with probability p, and 1 with probability $1-p$;
5: Generate a response indicator vector $r^{(3)}$ of length n such that its jth element is equal to $r_j^{(2)}$ with probability r, and r_j with probability $1-r$;
6: Generate missingness in $y^{(3)}$ according to the response indicator $r^{(3)}$.

then $\mathbb{C}\mathrm{or}(R_1, R_3) = \rho$.

PROOF. We have that

$$\begin{aligned}
\mathbb{C}\mathrm{ov}(R_1, R_3) &= \mathbb{E}(R_1 R_3) - \mathbb{E}(R_1)\mathbb{E}(R_3) \\
&= \mathbb{E}\left(I\left\{U \leqslant \rho\right\} R_1 + I\left\{U > \rho\right\} R_1 R_2\right) - \mathbb{E}(R_1)\mathbb{E}\left(I\left\{U \leqslant \rho\right\} R_1 + I\left\{U > \rho\right\} R_2\right) \\
&= \mathbb{P}\left\{U \leqslant \rho\right\}\mathbb{E}(R_1) + \mathbb{P}\left\{U > \rho\right\}\mathbb{E}(R_1)\mathbb{E}(R_2) - \mathbb{E}(R_1)\mathbb{E}(R_1)\mathbb{P}\left\{U \leqslant \rho\right\} - \mathbb{E}(R_1)\mathbb{E}(R_2)\mathbb{P}\left\{U > \rho\right\} \\
&= \rho q + (1-\rho)q^2 - \rho q^2 - (1-\rho)q^2 \\
&= \rho q(1-q).
\end{aligned}$$

Additionally,

$$\sqrt{\mathbb{V}\mathrm{ar}(R_1)} = \sqrt{q(1-q)}$$

and

$$\begin{aligned}
\mathbb{V}\mathrm{ar}(R_3) &= \mathbb{E}\left(R_3^2\right) - (\mathbb{E}(R_3))^2 \\
&= \mathbb{E}\left(I\left\{U \leqslant \rho\right\} R_1 + 2I\left\{U \leqslant \rho\right\} I\left\{U > \rho\right\} R_1 R_2 + I\left\{U > \rho\right\} R_2\right) - \left(rq + (1-r)q\right)^2 \\
&= rq + 2\mathbb{E}\left(I\left\{U \leqslant \rho\right\}\left(1 - I\left\{U \leqslant \rho\right\}\right)\right)q^2 + (1-\rho)q - q^2 \\
&= q - q^2 + 2 \cdot 0 \\
&= q(1-q),
\end{aligned}$$

so $\sqrt{\mathbb{V}\mathrm{ar}(R_3)} = \sqrt{q(1-q)}$. Finally, we have that

$$\mathbb{C}\mathrm{or}(R_1, R_3) = \frac{\mathbb{C}\mathrm{ov}(R_1, R_3)}{\sqrt{\mathbb{V}\mathrm{ar}(R_1)}\sqrt{\mathbb{V}\mathrm{ar}(R_3)}} = \frac{\rho q(1-q)}{\sqrt{q(1-q)}\sqrt{q(1-q)}} = \rho.$$

■

REMARK 3.7. Similarly to the missingness settings shown in Figures 3.5–3.7, we adapt Algorithm 3.1 so that the variable $Y^{(1)}$ governs the missingness of $Y^{(2)}$, while the missingness in $Y^{(3)}$ is generated to be correlated with that of $Y^{(2)}$. Additionally, MCAR is imposed on $Y^{(1)}$ to make the alternative hardly detectable for the A_n-based test.

3.4.2 Performance of the tests under zero mean or uncorrelated response indicators

In this subsection, we present the results of simulations in which the A_n'-based test was conducted under the assumption that the variables have zero means or that the response indicators are uncorrelated. In this case, no scaling of the data was performed prior to the test, as

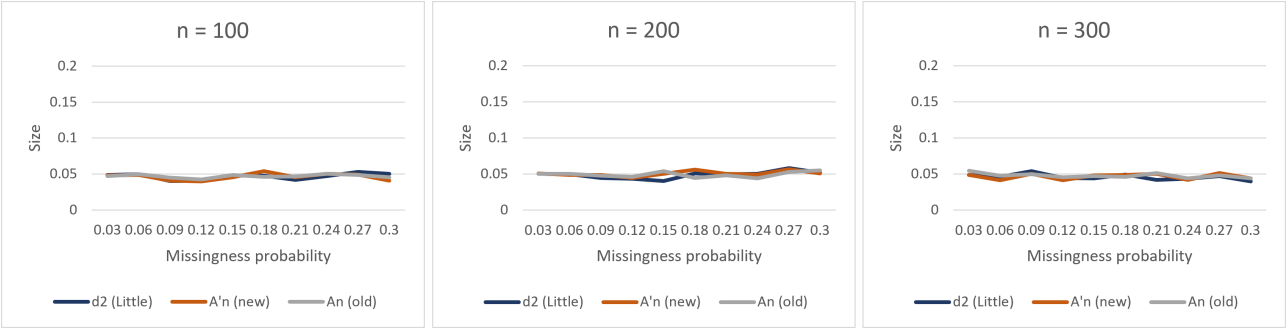

Figure 3.1: Empirical type I errors for 2X3Y case, standard normal distribution.

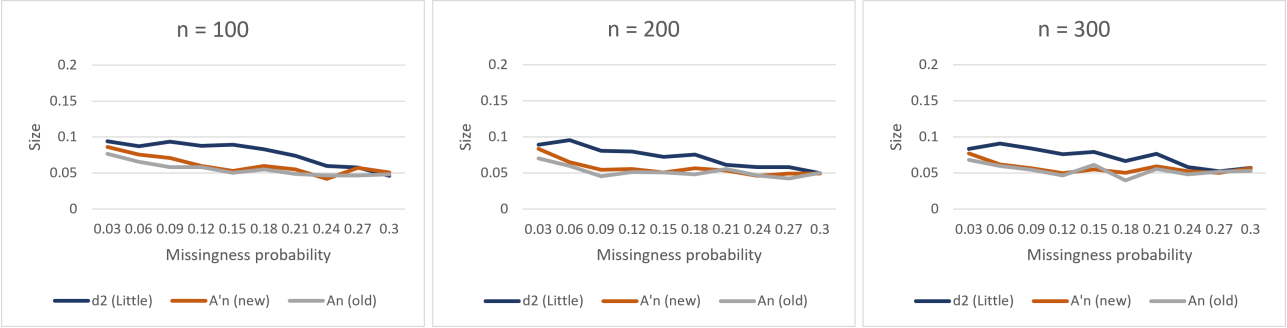

Figure 3.2: Empirical type I errors for 2X3Y case, Clayton copula with parameter 1 and $\mathcal{E}(1)$ margins.

it was unnecessary for its validity. We first present results for the exact missingness scenarios considered in Aleksić (2024), and subsequently for scenarios in which the alternative was undetectable or only marginally detectable by the original test. Finally, we discuss the case of increasing dimension.

Performance in scenarios where A_n-based test was compared to Little's test

As one can see from Figure 3.1, for the standard normal distribution, all three tests are well calibrated and have the empirical type I error approximately equal to the nominal level. From Figures 3.2 and 3.3 we can see that Little's d^2 has significantly larger deviation of the type I error, which is almost twice the nominal level. However, in most of the real-world scenarios that would not be the problem, especially since the empirical type I error remains stable across sample sizes. On the other hand, A_n and A'_n have very similar performance and are much better calibrated compared to d^2. This is most clearly seen in Figure 3.2, where the data distribution deviates most from normality.

Figure 3.4 shows that, under *MAR 1 to x* alternative and normal data, the novel test based on A'_n suffers a power loss compared to old one based on A_n, but it still outperforms Little's MCAR test, especially for smaller sample sizes. A similar conclusion holds for other underlying

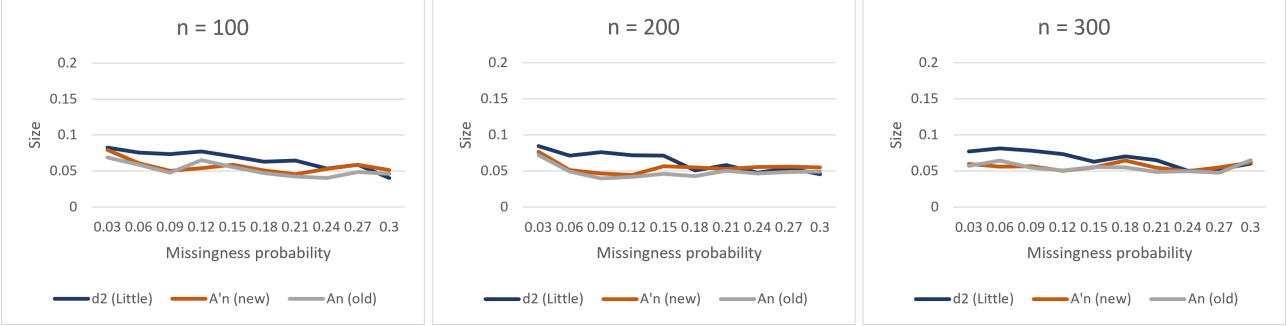

Figure 3.3: Empirical type I errors for 2X3Y case, Clayton copula with parameter 1 and χ^2_4 margins.

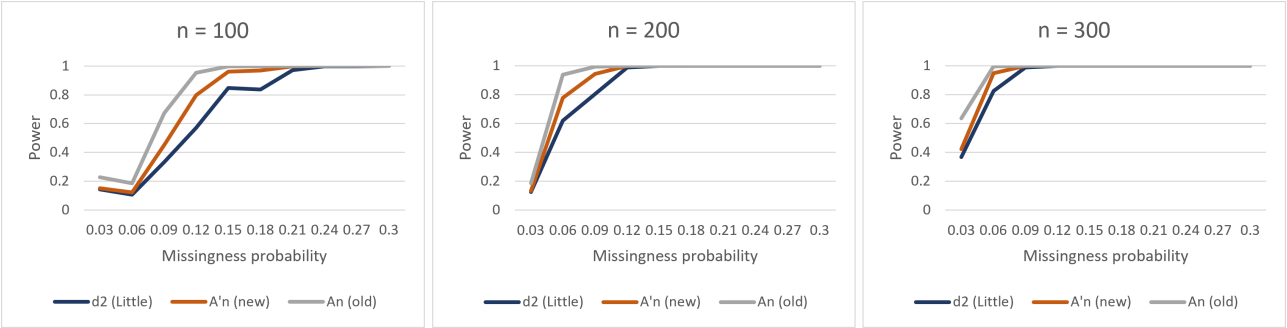

Figure 3.4: Empirical test powers for 2X3Y case, standard normal distribution, MAR 1 to 9 (var. 1 controls missingness in var. 3 and var. 5, var. 2 controls var. 4).

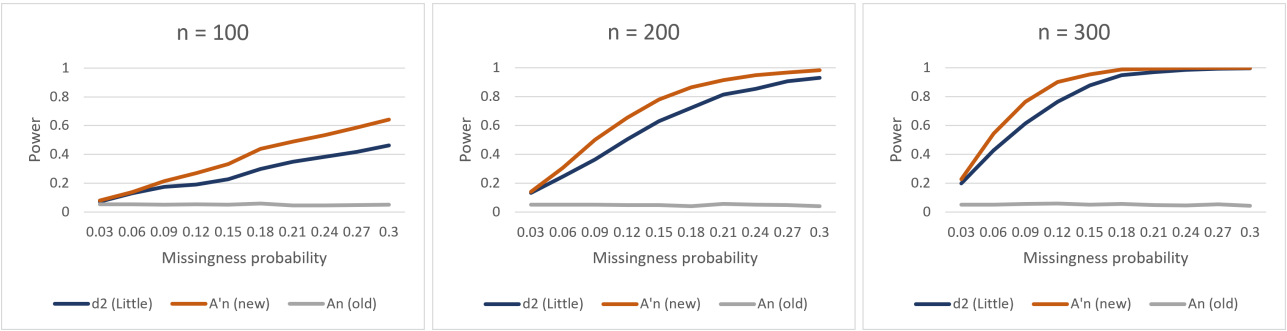

Figure 3.5: Empirical test powers for 2X3Y case, standard normal distribution, combination of MAR rank and MCAR (var. 3 controls missingness in var. 4 and var. 5, and then MCAR missingness is generated in var. 3).

distributions, as well as for the *MAR rank* mechanism. To improve the readability of the thesis, those can be found in the Supplementary Material of accompanying paper (Aleksić, 2025a).

Performance in novel scenarios

Figure 3.5 shows power performance for a specific MAR setting for standard normal $2X3Y$ data: variable 3 controls missingness in variables 4 and 5 according to *MAR rank* mechanism, and MCAR missingness is generated in variable 3 afterwards. This is a representative example of a setting where response indicators depend on the column which is incomplete - alternative undetectable for the old test. The novel test has once again performed better than Little's. Figure 3.6 shows that the old test is able to detect the alternative when the variables are correlated, namely the case of Clayton copula with parameter 1 and $\mathscr{E}(1)$ margins. The old test is able to capture the dependence through the completely observed ones. However, the old test has significantly lower power, whereas the novel test is comparable to Little's, although

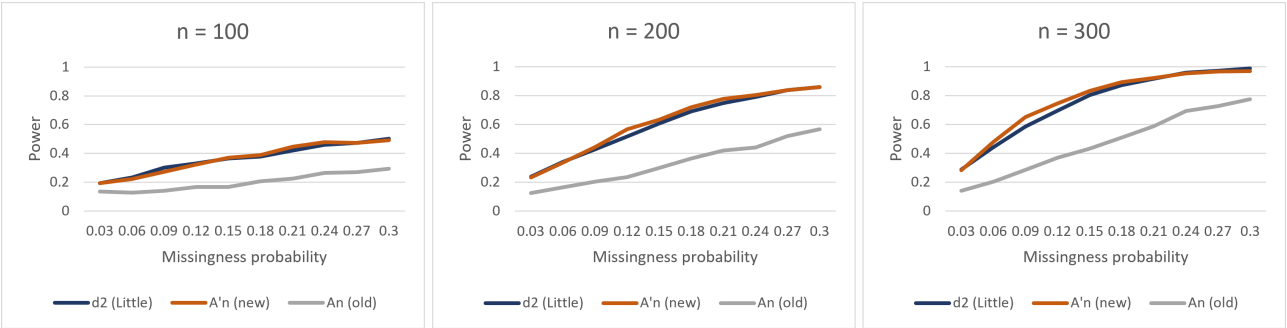

Figure 3.6: Empirical test powers for 2X3Y case, Clayton copula with $\mathscr{E}(1)$ margins, combination of MAR rank and MCAR (var. 3 controls missingness in var. 4 and var. 5, and then MCAR missingness is generated in var. 3).

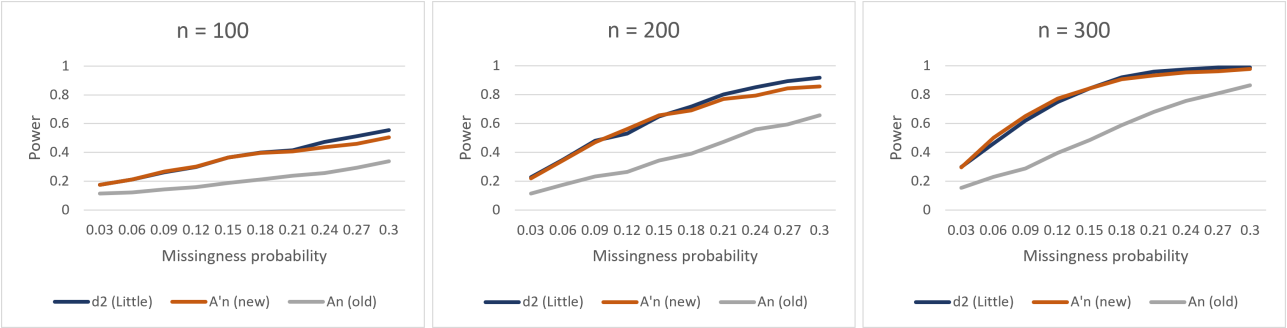

Figure 3.7: Empirical test powers for 2X3Y case, Clayton copula with χ^2_4 margins, combination of MAR rank and MCAR (var. 3 controls missingness in var. 4 and var. 5, and then MCAR missingness is generated in var. 3).

slightly more powerful. For χ^2_4 margins the old test is once again significantly less powerful than others, but we can see that Little's test has almost the same power as the novel one, having barely larger power for extremely large missingness rates that are not expected to be very common in practice. The behavior observed in Figures 3.5, 3.6, and 3.7 persists across different dimensions, distributions, and alternatives undetectable for A_n-based test: if the novel test has larger power than Little's, it is substantially better, and in other cases it is comparable, or slightly worse for large missingness rates. Further examples of this behavior are shown in the tables in the Supplementary Material of Aleksić (2025a).

REMARK 3.8. It is important to note that in scenarios where the data deviate from normality, Little's test tends to reject the null hypothesis more frequently than it should, indicating inflated type I error rates. Consequently, the apparent power of Little's test under non-normal settings should be interpreted with caution, as part of the observed rejections may come from this liberal behavior rather than genuine departures from the null. In other words, the reported powers for non-normal data likely overestimate the true power of the test. Therefore, when comparing the performance of the proposed methods, the results obtained under normality should be regarded as the most reliable benchmark, providing the best estimate of the actual differences in power between procedures.

Since all three studied tests rely on the assumption of all of the variables having finite fourth moments, it is interesting to examine the robustness of tests when that assumption is not fulfilled. Figure 3.8 shows the empirical type I errors for the standard Student's t-distribution with 2 degrees of freedom, which does not have finite fourth moments. As we can see, the novel test performs much better that Little's test, even for larger sample sizes. Despite the tendency of d^2-based test to reject the null hypothesis in this setting even when the null hypothesis is true, Figure 3.9 shows that the novel test is significantly more powerful. For a small number of variables (e.g., 3 or fewer), Little's test can be slightly more powerful in some scenarios, but that difference in power is notable for large missingness rates, which are not very common in practice. For clarity and better organization of the text, we omit those figures from the text.

Another important scenario in which the novel test needs to be examined is the case of MNAR data. Since the test by its construction is not able to calculate the covariance between the incomplete variable and its response indicators, alternatives that are "purely MNAR" should be undetectable for the test. More precisely, those are alternatives where the only form of dependence between the response indicators and the data is realized between the variable and its indicators, but not any others. However, as seen from Figure 3.10, in the case of the standard normal distribution, Little's test is not able to detect such alternative either. Figure 3.11 presents behavior in the case of same missingness mechanism, but Clayton copula with parameter 1 and exponential with parameter 1 margins. As we can see, all tests have practi-

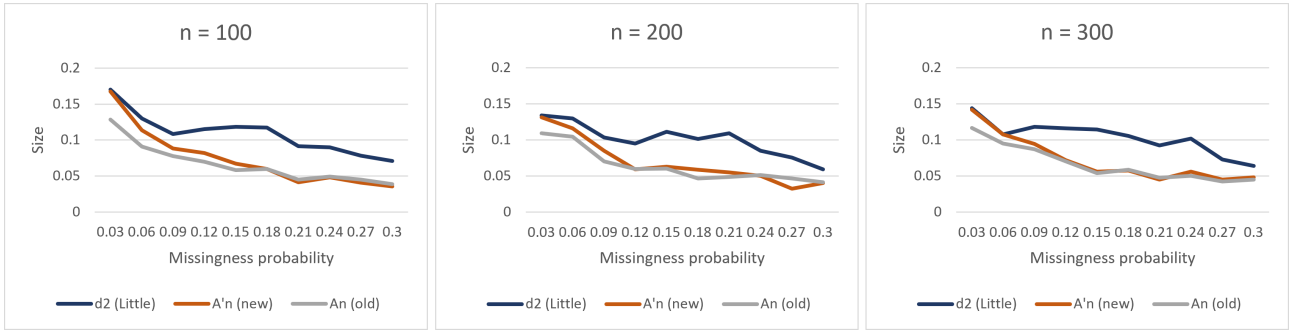

Figure 3.8: Empirical type I errors for 2X3Y case, standard Student's t_2 distribution.

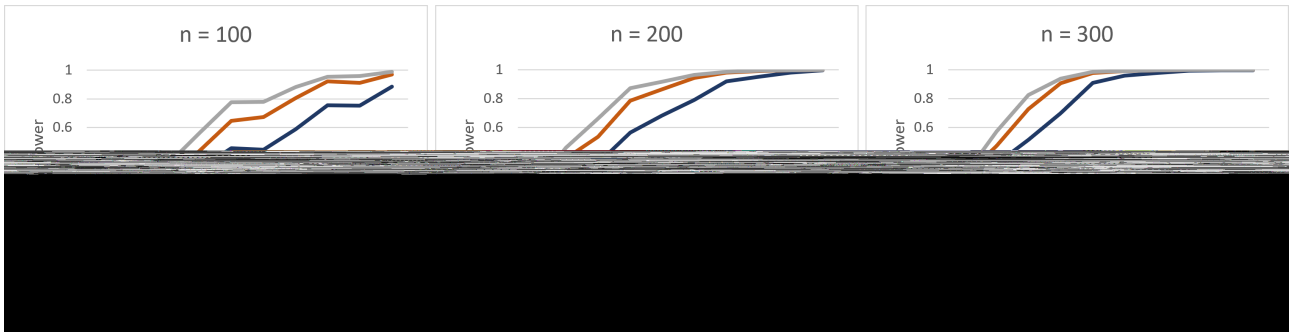

Figure 3.9: Empirical test powers for 2X3Y case, standard Student's t_2 distribution, MAR 1 to 9 (var. 1 controls missingness in var. 3 and var. 5, var. 2 controls var. 4).

cally the same power, and are able to detect the alternative. The same behavior is noted for other distributions and dimensions.

However, there are exceptions that behave unexpectedly, such as previously studied Student's t_2 distribution which does not have finite fourth moments. When combined with upper censoring as the only missingness mechanism, Figure 3.12 shows that all three test experience *loss* of power as the missingness rate increases, which is not expected, and was not observed in the scenarios before. It appears that the combination of an undetectable alternative and data from a population that do not satisfy the assumption of finite fourth moment is too challenging for the tests to handle, and they start behaving in a strange manner. We have tried replacing the identity scale matrix of the standard t_2 distribution with the matrix that has unit diagonal elements, and others equal to 0.1 and 0.5, respectively, but it did not help the Little's test, and behavior persisted. For example, for the scale matrix with non-diagonal elements equal to 0.5, the novel test stopped having decreasing power for $n = 100$, but Little's test stabilized for $n = 300$.

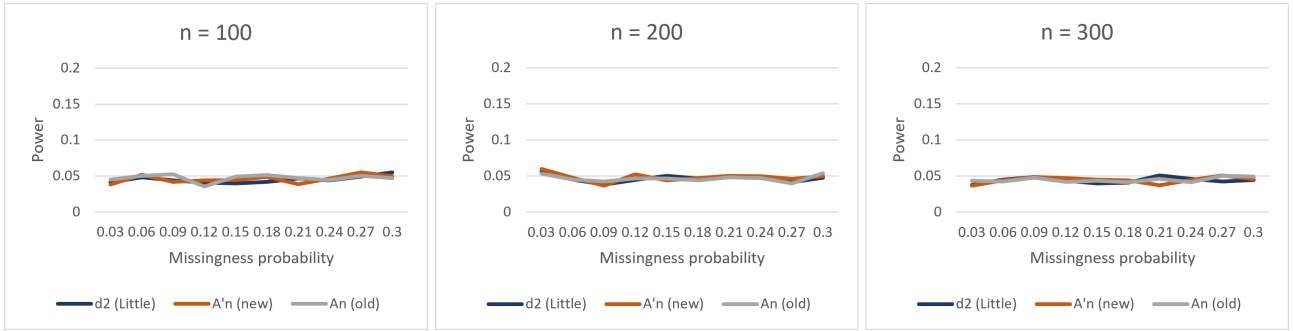

Figure 3.10: Empirical test powers for 2X3Y case, standard normal distribution, MNAR (upper) censoring.

Remark 3.9. The novel test we have presented is based on estimating the covariance between the response indicators and the data variables, which is a measure of linear dependence. To

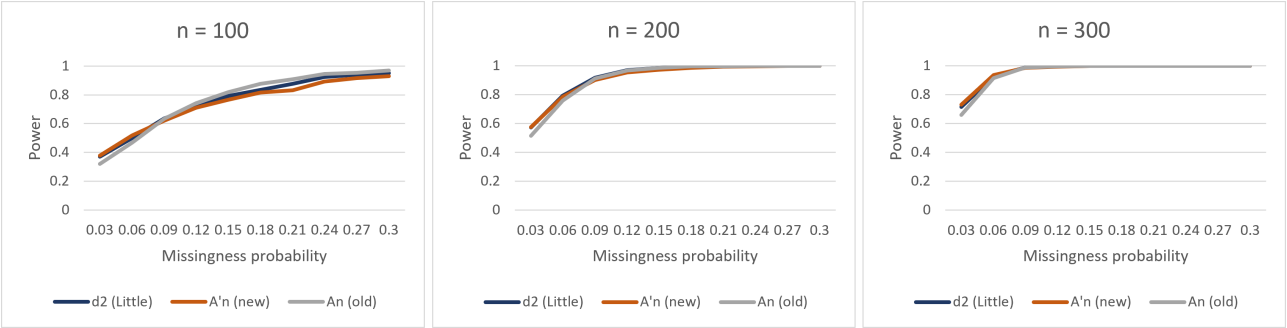

Figure 3.11: Empirical test powers for 2X3Y case, Clayton copula with parameter 1 and $\mathcal{E}(1)$ margins, MNAR (upper) censoring.

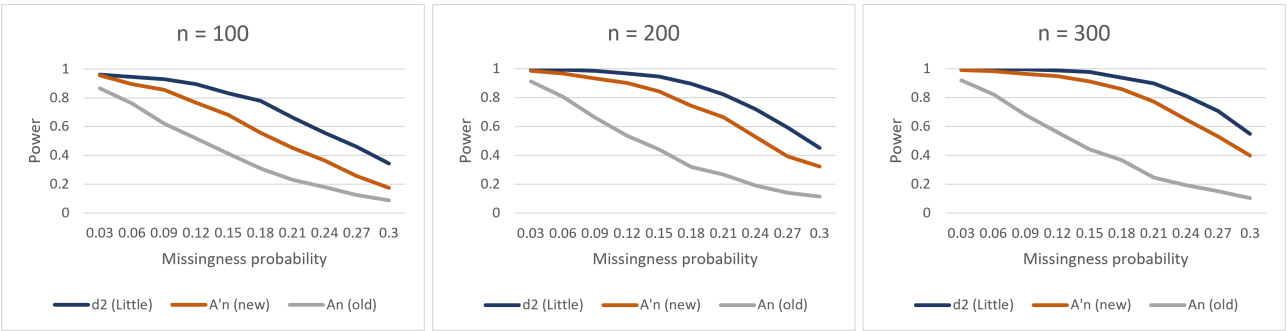

Figure 3.12: Empirical test powers for 2X3Y case, standard Student's t_2 distribution, MNAR (upper) censoring.

capture other form of dependence, one is free to transform the variables and apply the test to the transformed data, if some other form of dependence is expected to occur based on previous experience. One such example where transforming the variables improves the power performance of the test can be found in Aleksić (2024), where the original A_n-based test was introduced. The transformation of variables in this context could be the best solution in those cases where specific dependence between the data and the response indicators is to be expected.

3.4.3 Performance of the tests under nonzero mean and correlated response indicators

This subsection examines the properties of the A'_n-based test in situations where data centering is required, that is, when some variables have nonzero means and certain response indicators are correlated. We provide a concise overview of the conducted simulations to avoid overloading the text. For this purpose, only some of the results for the normal distribution with all means equal to 1 and covariance matrix $0.5I + 0.5J$ are presented. The behavior observed for other distributions is consistent with the results shown here, except that Little's test exhibits weaker type I error control for distributions that deviate substantially from normality. This was also the case for the setting of zero means and uncorrelated response indicators.

The same pattern is also observed across different dimensions and alternative hypotheses.

Figures 3.13 and 3.14 display the empirical type I errors for normally distributed data with correlated response indicators and nonzero means. As observed, all tests are well calibrated, while the A'_n-based test shows a slightly lower empirical type I error than the nominal 0.05 level, though the difference is minor and not practically significant. As previously noted, Little's MCAR test exhibits weaker control of the type I error when the data deviate from normality. Figure 3.15 provides an illustrative example of such behavior.

Figure 3.16 provides a representative example of the power behavior under an alternative detectable by the A_n-based test. As in the zero-mean case, the A'_n-based test outperforms

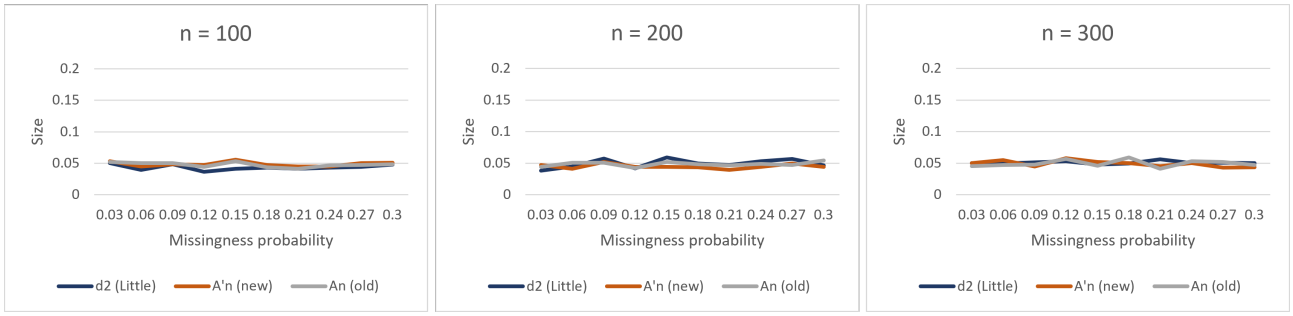

Figure 3.13: Empirical type I errors for 2X3Y case, normal distribution with mean $(1,1,1,1,1)$ and covariance matrix $0.5I_5 + 0.5J_5$, MCAR using Algorithm 3.1 with $\rho = 0.2$.

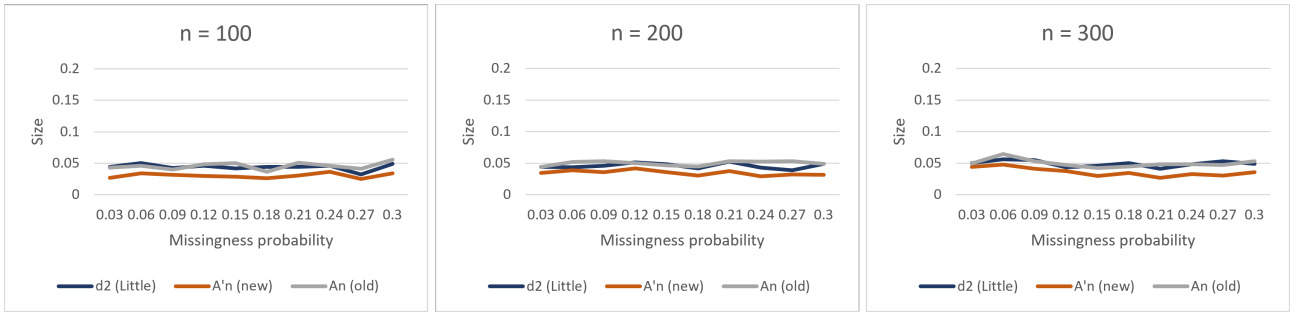

Figure 3.14: Empirical type I errors for 2X3Y case, normal distribution with mean $(1,1,1,1,1)$ and covariance matrix $0.5I_5 + 0.5J_5$, MCAR using Algorithm 3.1 with $\rho = 0.8$.

Little's test but is slightly outperformed by the A_n-based one. However, for the alternatives described in Remark 3.7, the A'_n-based test attains the highest power, particularly at higher missingness rates; a representative example is shown in Figure 3.17.

Behavior as dimensionality increases

Another important remark is that the standard implementation of Little's MCAR test in the `R` package `naniar` can handle no more than 30 variables. To the best of our knowledge, the most capable implementation addressing this limitation is found in the (now deprecated) `BaylorEdPsych` package, which can process up to 50 variables. In contrast, the novel test introduced here has no such constraints, neither theoretical nor practical.

To examine the type I error and power behavior, we conduct a series of simulations. Following our previous notation, we generate 2X3Y, 5X5Y, and 10X10Y datasets and compare the performance of the tests as dimensionality increases. MAR 1 to 9 data are generated as follows. For 2X3Y data, missingness is generated using Algorithm 3.1 with $r = 0.5$. For the 5X5Y case, the algorithm was modified so that the variables $X^{(1)}$, $X^{(2)}$, and $X^{(3)}$ govern the missingness in $Y^{(1)}$, $Y^{(2)}$, and $Y^{(3)}$, respectively. Subsequently, the missingness in $Y^{(4)}$ and $Y^{(5)}$ is generated to be correlated with that in $Y^{(2)}$ and $Y^{(3)}$, following the same procedure as in the original algorithm. For the 10X10Y case, the same procedure was applied, with variables $X^{(1)}$ through $X^{(5)}$ governing the missingness in $Y^{(1)}$ through $Y^{(5)}$. The variables $Y^{(6)}$ through $Y^{(10)}$ were then made incomplete, with their response indicators correlated to those of $Y^{(1)}$ through $Y^{(5)}$. MCAR data are generated in a similar manner, except that no variable pairing is used to govern missingness among variables.

Simulation results for a normal distribution with all variable means equal to 1, covariance matrix $0.5I + 0.5J$, and sample size $n = 100$, reveal that Little's test suffers a substantial loss of both type I error control and power as dimensionality increases. This behavior is illustrated in Figure 3.18 for type I error and in Figure 3.19 for power. As dimensionality grows, the empirical type I error of Little's test becomes much smaller than the nominal level, which

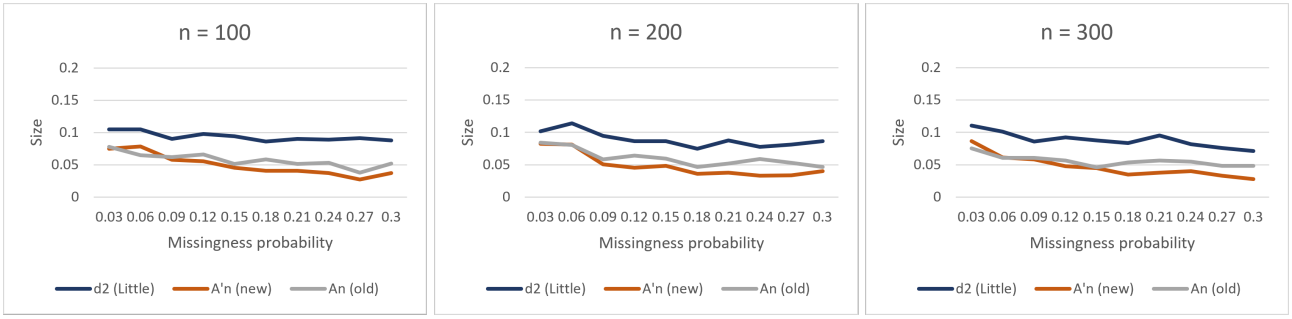

Figure 3.15: Empirical type I errors for 2X3Y case, Clayton copula with parameter 1 and $\mathscr{E}(1)$ margins, MCAR using Algorithm 3.1 with $\rho = 0.8$

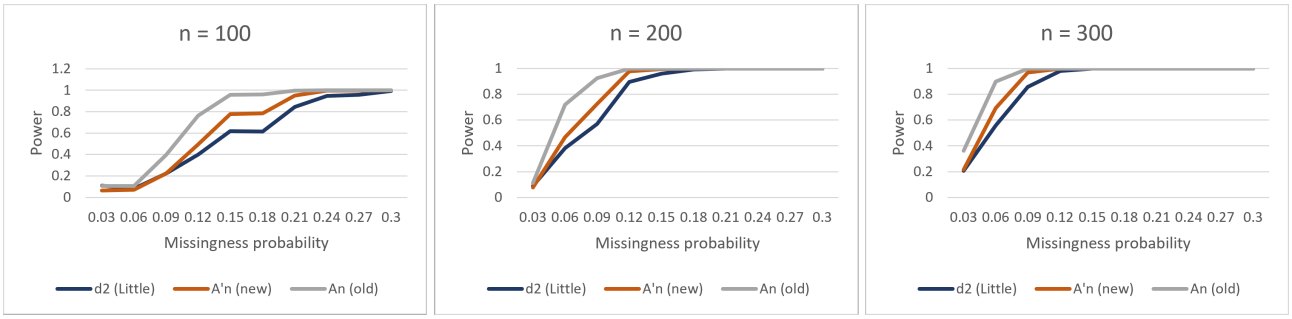

Figure 3.16: Empirical test powers for 2X3Y case, normal distribution with mean $(1,1,1,1,1)$ and covariance matrix $0.5I_5 + 0.5J_5$, MAR 1 to 9 using Algorithm 3.1 with $\rho = 0.8$.

negatively impacts its power.

In contrast, the A'_n-based test remains considerably more stable, with type I error showing noticeable inflation only for 20-dimensional data with high missingness rates; even then, the deviation is moderate. The A_n-based test is the most stable overall, with empirical type I error nearly equal to the nominal level. However, the improved A'_n-based test is preferable when no specific form of dependence among the response indicators can be assumed.

To evaluate the impact of using the more general A'_n-based test on power, we consider an A_n-detectable alternative. As shown in Figure 3.19, the generalized test exhibits lower power than the original A_n-based test, even in cases where type I error inflation occurs.

As shown in Figure 3.20, increasing the sample size to $n = 300$ allows the A'_n-based test to regain proper type I error control, with empirical values remaining close to the nominal level. In contrast, Little's test continues to exhibit the same issues observed for smaller samples: its empirical type I error remains well below the nominal value, resulting in a marked reduction in power relative to the other two tests.

The corresponding power performance is presented in Figure 3.21. The A_n-based test maintains consistent performance, while the A'_n-based test, though stable, has slightly lower power

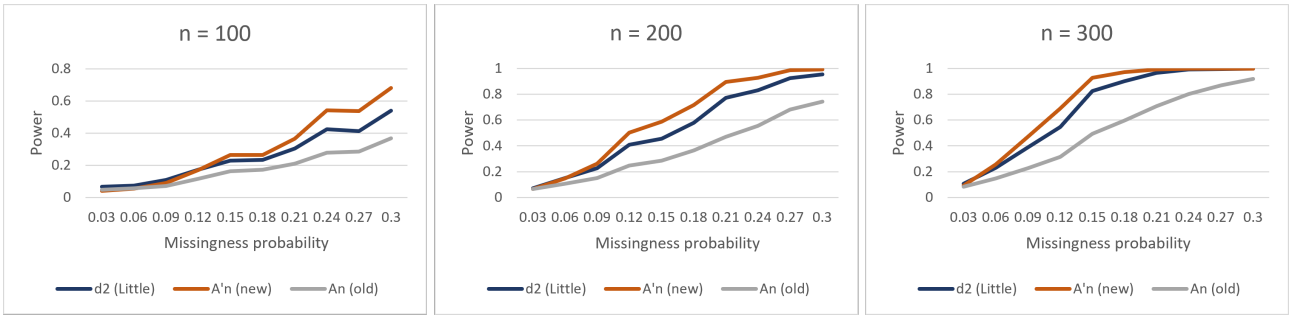

Figure 3.17: Empirical test powers for 2X3Y case, normal distribution with mean $(1,1,1,1,1)$ and covariance matrix $0.5I_5 + 0.5J_5$, MAR 1 to 9 using modified Algorithm 3.1 from Remark 3.7, with $\rho = 0.8$.

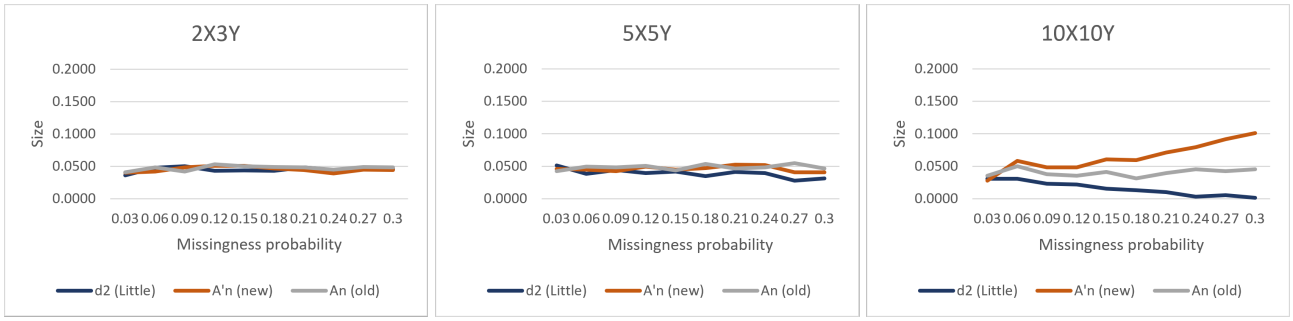

Figure 3.18: Empirical type I errors as dimension increases, MCAR with correlated response indicators, $\rho = 0.5$, $n = 100$.

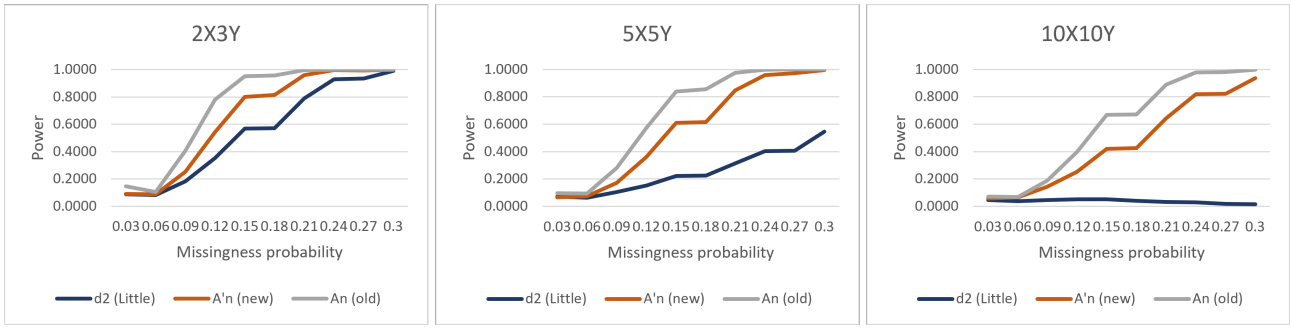

Figure 3.19: Empirical test powers as dimension increases, MAR 1 to 9 with correlated response indicators, $\rho = 0.5$, $n = 100$.

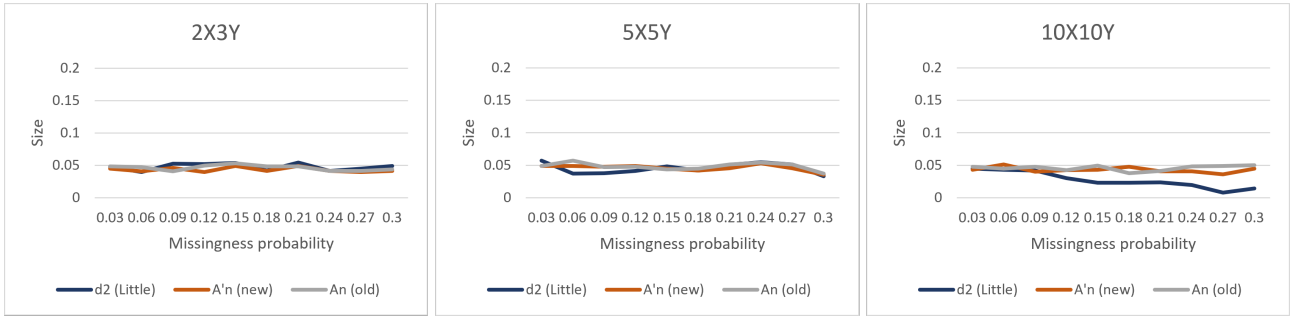

Figure 3.20: Empirical type I errors as dimension increases, MCAR with correlated response indicators, $\rho = 0.5$, $n = 300$.

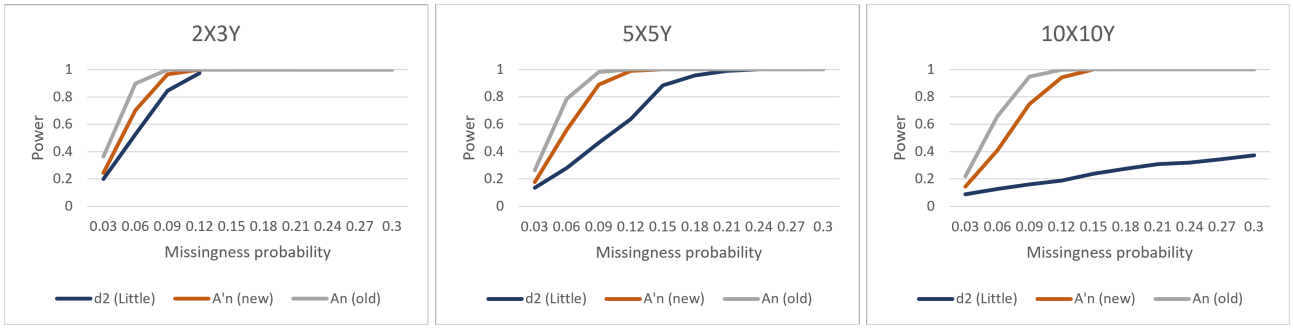

Figure 3.21: Empirical test powers as dimension increases, MAR 1 to 9 with correlated response indicators, $\rho = 0.5$, $n = 300$.

than the A_n-based one, similarly to the behavior observed in the smaller-sample case.

REMARK 3.10. We believe that the problems observed with Little's test come from how the test is constructed. It splits the data into groups based on missingness patterns and then estimates the mean and covariance within each group. When the number of variables is large and the total sample size is small, each group ends up with only a few observations, making these estimates very unreliable. This explains the poor control of type I error and loss of power in such settings.

Chapter 4

Non-degenerate U-statistics for MCAR data with application to testing independence

The main goal of this chapter is to derive the asymptotic distribution of a non-degenerate U-statistic under the MCAR assumption, presenting the results introduced by Aleksić et al. (2023). Sections 4.1 and 4.2 are devoted to the theoretical results. In Section 4.1 we derive the asymptotic distribution of a U-statistic with non-degenerate kernel under the *complete-case* approach. Section 4.2 applies these results results to the independence testing using Kendall's τ. Furthermore, as an alternative to the traditional *complete-case* approach, sample median imputation is considered, and the asymptotic properties of Kendall's statistic are examined under that approach of handling missing data. The main result of this section is the derivation of the asymptotic distribution of Kendall's estimate on the median-imputed dataset. In Section 4.3 we compare the complete-case approach for handling missing data and the sample median imputation in the context of Kendall's test. An extensive simulation study is conducted to compare the type I error and the power for samples of small and moderate size. Finally, in Section 4.4 we illustrate the methodology on a real-data example.

4.1 Asymptotic distribution under the complete-case approach

In this section, we obtain the asymptotic distribution of a non-degenerate U-statistic in the presence of MCAR data and the complete-case approach for handling missing data. These results form the basis for obtaining asymptotic properties under different imputation approaches, which we briefly discuss after the formulation and proof of the main result.

Let $X_1, \dots, X_n$ be a sample of IID d-dimensional random vectors. Suppose that for the purpose of estimating of an unknown parameter θ we consider a non-degenerate U-statistic

$$\hat{\theta}_n = \frac{1}{\binom{n}{2}} \sum_{1 \leqslant i < j \leqslant n} \phi(X_i, X_j), \tag{4.1}$$

let us denote $\phi(i,j) := \phi(X_i, X_j)$ for its kernel and assume that the kernel is square-integrable. Then, by Theorem 1.1, we know that the following relation holds:

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{D} \mathcal{N}(0, 4\sigma_1^2), \tag{4.2}$$

where $\sigma_1^2 = \mathbb{C}\mathrm{ov}(\phi(1,2), \phi(1,3))$.

Now suppose that some of the data are MCAR and that every row has an overall probability of being incomplete p and denote $q = 1-p$. If we denote by S_i an indicator that X_i is completely observed (i.e. $S_i = \prod_{j=1}^{d} R_i^{(j)}$), then the overall number of complete rows is equal to $\sum_{i=1}^{n} S_i$. It is clear that $q = \mathbb{E}(S_i)$ for all $i = 1, 2, \ldots, n$. Under the MCAR assumption, indicators S_i are independent of any X_i. Due to the elements of a sample being IID, indicators S_i are also mutually independent.

A standard approach is, of course, to calculate a statistic $\hat{\theta}_n$ only on the complete rows, which we sometimes refer to as *complete-case U-statistic*. It can be written as:

$$\hat{\hat{\theta}}_n := \frac{1}{\binom{\hat{n}}{2}} \sum_{1 \leqslant i < j \leqslant n} \phi(i,j) S_i S_j, \tag{4.3}$$

where $\hat{n} = \sum_{i=1}^{n} S_i$. Since using the complete-case approach for handling missing data is equivalent to working with a sample of smaller size, we expect $\hat{\hat{\theta}}_n - \theta$ to have the same asymptotic distribution as $\hat{\theta}_n - \theta$ when we use $\hat{n}$ as a normalizing constant. In other words, we expect $\sqrt{\hat{n}}(\hat{\hat{\theta}}_n - \theta)$ to have the same asymptotic distribution as $\sqrt{n}(\hat{\theta}_n - \theta)$. Note that the size of a truncated sample is now a random variable, unlike the original size n. This is formalized in Theorem 4.1. Despite being somewhat obvious, the rigorous proof is not trivial and, to the best of our knowledge, can not be found in the literature.

THEOREM 4.1 [ALEKSIĆ, CUPARIĆ, MILOŠEVIĆ (2023)]**.** *Let $X_1, \ldots, X_n$ be a sample of IID d-dimensional random vectors, and let $\hat{\hat{\theta}}_n$ be as defined in* (4.3)*. It holds that*

$$\sqrt{n}(\hat{\hat{\theta}}_n - \theta) \xrightarrow{D} \mathcal{N}\left(0, \frac{4\sigma_1^2}{q}\right) \tag{4.4}$$

and

$$\sqrt{\hat{n}}\left(\hat{\hat{\theta}}_n - \theta\right) \xrightarrow{D} \mathcal{N}\left(0, 4\sigma_1^2\right), \tag{4.5}$$

as $n \to \infty$.

PROOF. Let us consider the expanded sample

$$\begin{bmatrix} X_1 & S_1 \\ X_2 & S_2 \\ \vdots & \vdots \\ X_n & S_n \end{bmatrix}$$

instead of the original one. Using this sample, we define the symmetric kernel

$$\tilde{\phi}\left((x_i, s_i), (x_j, s_j)\right) := \phi\left(x_i, x_j\right) s_i s_j.$$

We denote $\tilde{\phi}(i,j) := \phi(X_i, X_j) S_i S_j$. Let us consider the following U-statistic:

$$T_n := \frac{1}{\binom{n}{2}} \sum_{1 \leqslant i < j \leqslant n} \tilde{\phi}\left((X_i, S_i), (X_j, S_j)\right) = \frac{1}{\binom{n}{2}} \sum_{1 \leqslant i < j \leqslant n} \phi(i,j) S_i S_j.$$

First, due to the MCAR assumption, it is easy to see that $\mathbb{E}\big(\tilde{\phi}(i,j)\big) = q^2\theta$. Now, we have

$$\begin{aligned} \tilde{\sigma}_1^2 &= \mathbb{C}\mathrm{ov}\Big(\tilde{\phi}(1,2), \tilde{\phi}(1,3)\Big) \\ &= \mathbb{E}\big(\tilde{\phi}(1,2)\tilde{\phi}(1,3)\big) - q^4\theta^2 \end{aligned}$$

$$
\begin{aligned}
&= \mathbb{E}\big(\phi(1,2)S_1S_2\phi(1,3)S_1S_3\big) - q^4\theta^2 \\
&= \Big\{ S_j \text{ are independent of all } X_j \Big\} \\
&= \mathbb{E}(S_1^2)\mathbb{E}(S_2)\mathbb{E}(S_3)\mathbb{E}\big(\phi(1,2)\phi(1,3)\big) - q^4\theta^2.
\end{aligned}
$$

Since, $S^2 = S$ and hence $\mathbb{E}(S^2) = \mathbb{E}(S) = q$ we have

$$
\tilde{\sigma}_1^2 = q^3\mathbb{E}\big(\phi(1,2)\phi(1,3)\big) - q^4\theta^2 = q^3\big(\sigma_1^2 + \theta^2\big) - q^4\theta^2,
$$

which is obviously strictly positive, so T_n is non-degenerate.

By Theorem 1.1, it holds that

$$
\sqrt{n}\big(T_n - q^2\theta\big) \xrightarrow{D} \mathcal{N}(0, 2^2(q^3\sigma_1^2 + (q^3 - q^4)\theta^2)),
$$

or, equivalently:

$$
\sqrt{n}\left(\frac{T_n}{q^2} - \theta\right) \xrightarrow{D} \mathcal{N}\left(0, \frac{4\sigma_1^2}{q} + 4\frac{1-q}{q}\theta^2\right).
$$

Since we have that

$$
\hat{\hat{\theta}}_n = \frac{\binom{n}{2}}{\binom{\hat{n}}{2}} T_n,
$$

we can calculate

$$
\sqrt{n}\big(\hat{\hat{\theta}}_n - \theta\big) = \sqrt{n}\left(\frac{T_n}{q^2} - \theta\right) - \sqrt{n}\,T_n\left(\frac{1}{q^2} - \frac{\binom{n}{2}}{\binom{\hat{n}}{2}}\right).
$$

As we have that,

$$
\sqrt{n}\,T_n\left(\frac{1}{q^2} - \frac{\binom{n}{2}}{\binom{\hat{n}}{2}}\right) = -\frac{T_n}{q^2\bar{S}_n(\frac{1}{n} - \bar{S}_n)}\sqrt{n}(\bar{S}_n^2 - q^2) - \frac{1}{\sqrt{n}}\frac{\bar{S}_n - q^2}{q^2\bar{S}_n(\bar{S}_n - \frac{1}{n})}, \tag{4.6}
$$

where $\bar{S}_n = \frac{1}{n}\sum_{i=1}^n S_i$, by noting that the second summand is $o_{\mathbb{P}}(1)$, the only thing left to derive is the limiting distribution of the first summand.

Applying the Law of Large Numbers for U-statistics and the Continuous Mapping Theorem (e.g., Koroljuk and Borovskich, 2010), we get

$$
\frac{T_n}{q^2\bar{S}_n(\frac{1}{n} - \bar{S}_n)} \xrightarrow{P} -\frac{\theta}{q^2}. \tag{4.7}
$$

Then, applying the Delta Method for the function $g(q) = q^2$ we obtain that

$$
\sqrt{n}(\bar{S}_n^2 - q^2) \xrightarrow{D} \mathcal{N}(0, 4q^3(1-q)). \tag{4.8}
$$

Finally, using Slutsky's theorem on (4.7) and the (4.8), and symmetry of the normal distribution, it yields to

$$
\frac{T_n}{q^2\bar{S}_n(\frac{1}{n} - \bar{S}_n)}\sqrt{n}(\bar{S}_n^2 - q^2) \xrightarrow{D} \mathcal{N}\left(0, 4\frac{1-q}{q}\theta^2\right),
$$

as $n \to \infty$, which gives us that, indeed, $\sqrt{n}(\hat{\hat{\theta}}_n - \theta)$ has a limiting normal distribution with mean 0.

Since we know the limiting distributions of $\sqrt{n}\left(T_n/q^2-\theta\right)$ and $\sqrt{n}\,T_n\left(1/q^2-\binom{n}{2}/\binom{\hat{n}}{2}\right)$, it is sufficient to show that

$$\mathbb{C}\mathrm{ov}\left(\sqrt{n}\left(\frac{T_n}{q^2}-\theta\right),\sqrt{n}\,T_n\left(\frac{1}{q^2}-\frac{\binom{n}{2}}{\binom{\hat{n}}{2}}\right)\right)$$

behaves as

$$-\mathbb{V}\mathrm{ar}\left(\sqrt{n}\,T_n\left(\frac{1}{q^2}-\frac{\binom{n}{2}}{\binom{\hat{n}}{2}}\right)\right),$$

as $n\to\infty$. Having (4.6), it remains to calculate the limiting covariance of

$$\sqrt{n}\left(\frac{T_n}{q^2}-\theta\right)\quad\text{and}\quad\frac{T_n}{q^2\bar{S}_n(1/n-\bar{S}_n)}\sqrt{n}(\bar{S}_n^2-q^2).$$

We have

$$\mathbb{C}\mathrm{ov}\left(\sqrt{n}\left(\frac{T_n}{q^2}-\theta\right),\frac{T_n}{q^2\bar{S}_n\left(\frac{1}{n}-\bar{S}_n\right)}\sqrt{n}\left(\bar{S}_n^2-q^2\right)\right)=\mathbb{E}\left(\sqrt{n}\left(\frac{T_n}{q^2}-\theta\right)\frac{T_n}{q^2\bar{S}_n\left(\frac{1}{n}-\bar{S}_n\right)}\sqrt{n}\left(\bar{S}_n^2-q^2\right)\right).$$

For brevity, let us denote

$$\gamma_n=\frac{\bar{S}_n^2-q^2}{\bar{S}_n(\frac{1}{n}-\bar{S}_n)}.$$

Now, since

$$T_n^2=\frac{1}{\binom{n}{2}^2}\left(\sum_{i<j}\phi(i,j)S_iS_j\right)^2=\frac{1}{\binom{n}{2}^2}\left(\frac{1}{2}\sum_{i\neq j}\phi(i,j)S_iS_j\right)^2,$$

it holds that

$$\mathbb{E}\left(T_n^2\gamma_n\right)=\frac{1}{4\binom{n}{2}^2}\sum_{i\neq j}\sum_{k\neq l}\mathbb{E}\left(\phi(i,j)\phi(k,l)\right)\mathbb{E}\left(S_iS_jS_kS_l\gamma_n\right).\tag{4.9}$$

Next, we calculate the following sums, the importance of which will become obvious very soon:

$$\begin{aligned}
K_{1213}&=\sum_{i=1}^{n}\sum_{\substack{j=1\\j\neq i}}^{n}\sum_{\substack{l=1\\l\neq i\\l\neq j}}^{n}S_iS_jS_l\\
&=\sum_{i=1}^{n}S_i\sum_{\substack{j=1\\j\neq i}}^{n}S_j\left(n\bar{S}_n-S_i-S_j\right)\\
&=\left(n\bar{S}_n-2\right)\sum_{i=1}^{n}S_i\left(n\bar{S}_n-S_i\right)\\
&=n^3\bar{S}_n\left(\bar{S}_n-\frac{1}{n}\right)\left(\bar{S}_n-\frac{2}{n}\right),\\
K_{1212}&=\sum_{i=1}^{n}\sum_{\substack{j=1\\j\neq i}}^{n}S_iS_j=n^2\bar{S}_n\left(\bar{S}_n-\frac{1}{n}\right),
\end{aligned}$$

$$K_{1234}=\sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n}\sum_{\substack{k=1\\ k\neq i\\ k\neq j}}^{n}\sum_{\substack{l=1\\ l\neq i\\ l\neq j\\ l\neq k}}^{n}S_iS_jS_kS_l=n^4\bar{S}_n\left(\bar{S}_n-\frac{1}{n}\right)\left(\bar{S}_n-\frac{2}{n}\right)\left(\bar{S}_n-\frac{3}{n}\right).$$

Now, if we denote $\sigma_2^2=\mathbb{C}\mathrm{ov}(\phi(1,2),\phi(1,2))$, we have that:

$$\begin{aligned}
&\frac{n}{q^4}\mathbb{E}\left(T_n^2\frac{\bar{S}_n^2-q^2}{\bar{S}_n\left(\frac{1}{n}-\bar{S}_n\right)}\right)\\
&\quad=\frac{n}{q^4}\frac{1}{4\binom{n}{2}^2}\left(\sum_{i\neq j}\sum_{k\neq l}\mathbb{E}\big(\phi(i,j)\phi(k,l)\big)\mathbb{E}\big(S_iS_jS_kS_l\gamma_n\big)\right)\\
&\quad=\frac{n}{q^4}\frac{1}{4\binom{n}{2}^2}\left(\sum_{\substack{\text{all 4}\\ \text{indices}\\ \text{diff.}}}\theta^2\mathbb{E}(S_iS_jS_kS_l\gamma_n)+\sum_{\substack{\text{1 pair}\\ \text{the same}\\ \text{1 diff.}}}(\sigma_1^2+\theta^2)\mathbb{E}(S_iS_jS_kS_l\gamma_n)\right.\\
&\qquad\qquad\left.+\sum_{\substack{\text{both}\\ \text{pairs}\\ \text{the same}}}(\sigma_2^2+\theta^2)\mathbb{E}(S_iS_jS_kS_l\gamma_n)\right)\\
&\quad=\frac{n}{q^4}\frac{1}{4\binom{n}{2}^2}\big(\theta^2\mathbb{E}\big(\gamma_nK_{1234}\big)+4(\sigma_1^2+\theta^2)\mathbb{E}\big(\gamma_nK_{1213}\big)+(\sigma_2^2+\theta^2)\mathbb{E}(\gamma_nK_{1212})\big)\\
&\quad=\frac{1}{q^4}\frac{1}{n(n-1)^2}\theta^2\mathbb{E}\big(\gamma_nK_{1234}\big)+\frac{1}{q^4}\frac{1}{n(n-1)^2}\Big(4(\sigma_1^2+\theta^2)\mathbb{E}\big(\gamma_nK_{1213}\big)+(\sigma_2^2+\theta^2)\mathbb{E}(\gamma_nK_{1212})\Big).
\end{aligned}$$

As n approaches infinity, $(\sigma_2^2+\theta^2)\mathbb{E}(\gamma_nK_{1212})$ behaves as n^2. Therefore the last summand above vanishes as $n\to\infty$. Next,

$$\begin{aligned}
\mathbb{E}\big(\gamma_nK_{1213}\big)&=\mathbb{E}\left(\frac{\bar{S}_n^2-q^2}{\bar{S}_n\left(\frac{1}{n}-\bar{S}_n\right)}n^3\bar{S}_n\left(\bar{S}_n-\frac{1}{n}\right)\left(\bar{S}_n-\frac{2}{n}\right)\right)\\
&=-n^3\mathbb{E}\left(\left(\bar{S}_n^2-q^2\right)\left(\bar{S}_n-\frac{2}{n}\right)\right)\\
&=-\mathbb{E}\big(n^3\bar{S}_n^3\big)+n^3q^3+n^2\mathbb{E}\big(2\big(\bar{S}_n^2-q^2\big)\big).
\end{aligned}$$

Since

$$\mathbb{E}\big(n^3\bar{S}_n^3\big)=\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}\mathbb{E}(S_iS_jS_k)=n(n-1)(n-2)q^3+3n(n-1)q^2+nq,$$

we have that

$$\mathbb{E}\big(\gamma_nK_{1213}\big)=-n(n-1)(n-2)q^3-3n(n-1)q^2-nq+n^3q^3+n^2\mathbb{E}\big(2\big(\bar{S}_n^2-q^2\big)\big),$$

from which we conclude that terms having n^3 cancel out, so $\mathbb{E}\big(\gamma_nK_{1213}\big)$ is of order n^2. That leads us to the conclusion that

$$\frac{n}{q^4}\mathbb{E}\left(T_n^2\frac{\bar{S}_n^2-q^2}{\bar{S}_n\left(\frac{1}{n}-\bar{S}_n\right)}\right)=\frac{1}{q^4}\frac{1}{n(n-1)^2}\theta^2\mathbb{E}\big(\gamma_nK_{1234}\big)+u_n,$$

where $u_n \to 0$ as $n \to \infty$. Now,

$$\begin{aligned}
\mathbb{E}\left(\gamma_n K_{1234}\right) &= \mathbb{E}\left(\frac{\bar{S}_n^2 - q^2}{\bar{S}_n\left(\frac{1}{n} - \bar{S}_n\right)} n^4 \bar{S}_n \left(\bar{S}_n - \frac{1}{n}\right)\left(\bar{S}_n - \frac{2}{n}\right)\left(\bar{S}_n - \frac{3}{n}\right)\right) \\
&= -\mathbb{E}\left(n^4 \bar{S}_n^4\right) + 5\mathbb{E}\left(n^3 \bar{S}_n^3\right) + n^4 q^2 \mathbb{E}\left(\left(\bar{S}_n^2\right) - 5n^3 q^2 \mathbb{E}\left(\bar{S}_n\right) + O(n^2)\right. \\
&= -n(n-1)(n-2)(n-3)q^4 - 6n(n-1)(n-2)q^3 - 7n(n-1)q^2 - nq \\
&\quad + 5n(n-1)(n-2)q^3 + 15n(n-1)q^2 + 5nq + n^4 q^2\left(\frac{q(1-q)}{n} + q^2\right) - 5n^3 q^3 + O(n^2) \\
&= 6n^3 q^4 - n(n-1)(n-2)q^3 - 4n^3 q^3 - n^3 q^4 + O(n^2).
\end{aligned}$$

From here, we can easily conclude that

$$\begin{aligned}
\lim_{n\to\infty}\left(\frac{n}{q^4}\mathbb{E}\left(T_n^2 \frac{\bar{S}_n^2 - q^2}{\bar{S}_n\left(\frac{1}{n} - \bar{S}_n\right)}\right)\right) &= \lim_{n\to\infty}\left(\frac{1}{q^4}\frac{1}{n(n-1)^2}\theta^2 \mathbb{E}\left(\gamma_n K_{1234}\right) + u_n\right) \\
&= \frac{\theta^2}{q^4}\left(6q^4 - q^3 - 4q^3 - q^4\right) \\
&= -5\frac{1-q}{q}\theta^2.
\end{aligned}$$

Now let us go back and finish the calculation of (4.9):

$$\begin{aligned}
\mathbb{E}\left(T_n \frac{\bar{S}_n^2 - q^2}{\bar{S}_n\left(\frac{1}{n} - \bar{S}_n\right)}\right) &= \mathbb{E}\left(\frac{1}{\binom{n}{2}}\sum_{i<j}\phi(i,j)S_i S_j \frac{\bar{S}_n^2 - q^2}{\bar{S}_n\left(\frac{1}{n} - \bar{S}_n\right)}\right) = \frac{\theta}{\binom{n}{2}}\sum_{i<j}\mathbb{E}\left(S_i S_j \frac{\bar{S}_n^2 - q^2}{\bar{S}_n\left(\frac{1}{n} - \bar{S}_n\right)}\right) \\
&= \frac{\theta}{\binom{n}{2}}\mathbb{E}\left(\frac{\bar{S}_n^2 - q^2}{\bar{S}_n\left(\frac{1}{n} - \bar{S}_n\right)}\sum_{i<j} S_i S_j\right) = \frac{\theta}{n-1}\mathbb{E}\left(\frac{\bar{S}_n^2 - q^2}{\frac{1}{n} - \bar{S}_n} n\left(\bar{S}_n - \frac{1}{n}\right)\right) \\
&= -\frac{\theta}{n-1}q(1-q),
\end{aligned}$$

and hence we obtain that

$$-\frac{n\theta}{q^2}\mathbb{E}\left(T_n \frac{\bar{S}_n^2 - q^2}{\bar{S}_n\left(\frac{1}{n} - \bar{S}_n\right)}\right) = \frac{n}{n-1}\theta^2\frac{1-q}{q} \to \theta^2\frac{1-q}{q},$$

as $n \to \infty$.

Finally, we can conclude that

$$\lim_{n\to\infty}\mathbb{C}\mathrm{ov}\left(\sqrt{n}\left(\frac{T_n}{q^2} - \theta\right), \frac{T_n}{q^2\bar{S}_n\left(\frac{1}{n} - \bar{S}_n\right)}\sqrt{n}\left(\bar{S}_n^2 - q^2\right)\right) = -4\frac{1-q}{q}\theta^2,$$

which, together with the established asymptotic normality with mean zero and the fact that $\mathbb{V}\mathrm{ar}(X+Y) = \mathbb{V}\mathrm{ar}(X) + \mathbb{V}\mathrm{ar}(Y) + 2\mathbb{C}\mathrm{ov}(X,Y)$, finishes the proof of (4.4). Another application of Slutsky's theorem gives us (4.5), which completes the proof of Theorem 4.1. ■

If the decision is made not to proceed with the complete-case analysis, but to impute the data using some simple imputation methods like mean and median imputation, test statistics can be represented as a U-statistic with estimated parameters. For the asymptotic properties of such statistics we refer to the results discussed in Subsections 1.5.1 and 1.5.2, i.e. the papers by Randles (1982), De Wet and Randles (1987), and Cuparić et al. (2022). Applying results

therein, by eventually posing some additional mild conditions, the asymptotic properties for such statistics might be obtained. In the next section, we demonstrate this methodology by deriving the asymptotic distribution of Kendall's test statistic under the null hypothesis of independence, assuming MCAR data with uncorrelated response indicators and using the sample median imputation approach as method for handling missing data.

4.2 Testing independence using Kendall's tau

In this section, we examine the problem of testing independence for MCAR data using Kendall's coefficient, which quantifies the strength and direction of the association between two numerical random variables. We explore the testing using two approaches for handling missing data: the complete-case approach with results from Theorem 4.1, and the median-based imputation approach with theoretical results, where we derive the asymptotic distribution of Kendall's tau statistic computed from the median-imputed data.

Let (X, Y) be a two-dimensional random vector. Kendall's tau rank correlation coefficient between the variables X and Y is defined as $\mathbb{E}\big(\operatorname{sgn}(X-\tilde{X})\operatorname{sgn}(Y-\tilde{Y})\big)$, where $(\tilde{X}, \tilde{Y})$ is an independent copy of (X, Y). Given a sample of IID random vectors $(X_1, Y_1), \ldots, (X_n, Y_n)$, an unbiased estimator of τ is given by

$$\hat{\tau}_n = \frac{1}{\binom{n}{2}} \sum_{1 \leqslant i < j \leqslant n} \operatorname{sgn}(X_i - X_j)\operatorname{sgn}(Y_i - Y_j)$$

Note that this is a U- statistic with a non-degenerate kernel $\phi\big((x_1, y_1), (x_2, y_2)\big) = \operatorname{sgn}(x_1 - x_2)\operatorname{sgn}(y_1 - y_2)$, from which easily follows that

$$\sqrt{n}(\hat{\tau}_n - \tau) \xrightarrow{D} \mathcal{N}(0, \sigma_\tau^2),$$

where

$$\sigma_\tau^2 = 4\mathbb{V}\mathrm{ar}\big(\mathbb{E}\big(\operatorname{sgn}(X-\tilde{X})\operatorname{sgn}(Y-\tilde{Y}) \mid X, Y\big)\big).$$

Considering this, one may construct test for the independence of X and Y using $\hat{\tau}_n$ as the test statistic, and to reject the null hypothesis if $|\hat{\tau}_n|$ is sufficiently large. If data are MCAR , Theorem 4.1 tells us that we can use $\hat{\tilde{\tau}}_n$, defined as

$$\hat{\tilde{\tau}}_n = \frac{1}{\binom{\hat{n}}{2}} \sum_{1 \leqslant i < j \leqslant n} \operatorname{sgn}(X_i - X_j)\operatorname{sgn}(Y_i - Y_j) S_i S_j,$$

as the test statistic and preserve the asymptotic properties.

Moreover, under null hypothesis the asymptotic variance σ_τ^2 does not depend on the data distribution and is known to be equal to $4/9$ (see e.g., Kendall, 1975, p. 71).

4.2.1 Median imputation

In this subsection, we explore the limiting properties of an estimator of Kendall's tau based on median-imputed dataset.

Let M_X denote the median value obtained from the non-missing values among $X_1, \ldots, X_n$, and let M_Y be defined in a similar manner. Consider the following statistic:

$$\tilde{\tilde{\tau}}_n = \frac{1}{\binom{n}{2}} \sum_{1 \leqslant i < j \leqslant n} \Big(\operatorname{sgn}(X_i - X_j)\operatorname{sgn}(Y_i - Y_j) R_i^X R_j^X R_i^Y R_j^Y$$

$$
\begin{aligned}
&+\operatorname{sgn}(X_i-M_X)\operatorname{sgn}(Y_i-Y_j)R_i^X(1-R_j^X)R_i^Y R_j^Y\\
&+\operatorname{sgn}(M_X-X_j)\operatorname{sgn}(Y_i-Y_j)(1-R_i^X)R_j^X R_i^Y R_j^Y\\
&+\operatorname{sgn}(X_i-X_j)\operatorname{sgn}(Y_i-M_Y)R_i^X R_j^X R_i^Y(1-R_j^Y)\\
&+\operatorname{sgn}(X_i-X_j)\operatorname{sgn}(M_Y-Y_j)R_i^X R_j^X(1-R_i^Y)R_j^Y\\
&+\operatorname{sgn}(X_i-M_X)\operatorname{sgn}(Y_i-M_Y)R_i^X(1-R_j^X)R_i^Y(1-R_j^Y)\\
&+\operatorname{sgn}(M_X-X_j)\operatorname{sgn}(Y_i-M_Y)(1-R_i^X)R_j^X R_i^Y(1-R_j^Y)\\
&+\operatorname{sgn}(X_i-M_X)\operatorname{sgn}(M_Y-Y_j)R_i^X(1-R_j^X)(1-R_i^Y)R_j^Y\\
&+\operatorname{sgn}(M_X-X_j)\operatorname{sgn}(M_Y-Y_j)(1-R_i^X)R_j^X(1-R_i^Y)R_j^Y\Big), \qquad (4.10)
\end{aligned}
$$

where we denoted by R_i^X an indicator that X_i is observed, and similarly by R_i^Y an indicator that Y_i is observed. In other words, the statistic $\tilde{\tilde{\tau}}_n$ is obtained by replacing every missing value with the corresponding marginal sample median.

One can note that $\tilde{\tilde{\tau}}_n$ itself is a U-statistic with estimated parameters, having a symmetric kernel as in the expression (4.10). Let us denote it as $\Phi\big((X_i,Y_i,R_i^X,R_i^Y),(X_j,Y_j,R_j^X,R_j^Y);(M_X,M_Y)\big)$.

Suppose that the data were MCAR, the response indicators are uncorrelated, and that the columns have proportions of observed values equal to q_1 and q_2, respectively. More precisely, let R_i^X be independent from R_j^Y for every i and let $\mathbb{E}(R_i^X)=q_1$, $\mathbb{E}(R_i^Y)=q_2$. The next theorem provides the asymptotic distribution of $\tilde{\tilde{\tau}}_n$ under the null hypothesis of independence of X and Y.

Theorem 4.2 [Aleksić, Cuparić, Milošević (2023)]**.** *Let $(X_1,Y_1),\dots,(X_n,Y_n)$ be a sample of IID absolutely continuous 2-dimensional random vectors and let $\tilde{\tilde{\tau}}_n$ be as defined in* (4.10)*. Suppose that, if f_X and f_Y are the corresponding marginal densities, they are bounded in some neighborhood of median of X and median of Y, respectively. If X_j and Y_j are independent for $j=1,\dots,n$, and q_1 and q_2 defined as before, it holds that, as $n\to\infty$,*

$$
\sqrt{n}\tilde{\tilde{\tau}}_n \xrightarrow{D} \mathcal{N}\left(0,\frac{4}{9}q_1q_2(3-3q_1+q_1^2)(3-3q_2+q_2^2)\right).
$$

Proof. The proof consists of two steps. In the first step, we derive the asymptotic distribution of $\sqrt{n}\tilde{\tau}_n$, where $\tilde{\tau}_n$ is defined by replacing M_X and M_Y in (4.10) with theoretical medians m_X and m_Y, respectively. In the second step, we show that $\sqrt{n}\tilde{\tilde{\tau}}_n$ and $\sqrt{n}\tilde{\tau}_n$ have the same asymptotic distribution.

First step. It can be easily shown that $\mathbb{E}(\tilde{\tau}_n)=0$. In addition, one can notice that $\tilde{\tau}_n$ itself is a U-statistic, having a symmetric kernel whose first projection is equal to:

$$
\begin{aligned}
&\varphi_1(x,y,r^x,r^y;m_X,m_Y)\\
&\quad=\mathbb{E}\left(\Phi\big((X_1,Y_1,R_1^X,R_1^Y),(X_2,Y_2,R_2^X,R_2^Y);(m_X,m_Y)\big)\,\middle|\,X_1=x,Y_1=y,R_1^X=r^X,R_1^Y=r^Y\right)\\
&\quad=\mathbb{E}\big(\operatorname{sgn}(x-X_2)\big)\mathbb{E}\big(\operatorname{sgn}(y-Y_2)\big)r^X r^Y q_1q_2\\
&\qquad\quad+\operatorname{sgn}(x-m_X)\mathbb{E}\big(y-Y_2\big)r^X(1-q_1)r^Y q_2\\
&\qquad\quad+\mathbb{E}\big(\operatorname{sgn}(x-X_2)\big)\operatorname{sgn}(y-m_Y)r^X q_1 r^Y(1-q_2)\\
&\qquad\quad+\operatorname{sgn}(x-m_X)\operatorname{sgn}(y-m_Y)r^X r^Y(1-q_1)(1-q_2)\\
&\quad=(2F_X(x)-1)(2F_Y(y)-1)r^X r^Y q_1q_2\\
&\qquad\quad+\operatorname{sgn}(x-m_X)(2F_Y(y)-1)r^X r^Y(1-q_1)q_2\\
&\qquad\quad+(2F_X(x)-1)\operatorname{sgn}(y-m_Y)r^X r^Y q_1(1-q_2)
\end{aligned}
$$

$$+\operatorname{sgn}(x-m_X)\operatorname{sgn}(y-m_Y)r^X r^Y(1-q_1)(1-q_2).$$

Next, note that since X is absolutely continuous, $F_X(X)$ follows a uniform distribution on the interval $[0,1]$. As the CDF of any absolutely continuous distribution is strictly increasing on its support, it holds that $\operatorname{sgn}(x-m_X)=\operatorname{sgn}(F_X(x)-\frac{1}{2})$ and $\operatorname{sgn}(y-m_Y)=\operatorname{sgn}(F_Y(y)-\frac{1}{2})$. Therefore, we can, without loss of generality, assume in further calculations that X and Y are independent and uniformly distributed on $[0,1]$. Thus, we can write φ_1 as:

$$\begin{aligned}\varphi_1(x,y,r^X,r^Y;m_X,m_Y)=&(2x-1)(2y-1)r^X r^Y q_1 q_2\\&+\operatorname{sgn}\left(x-\frac{1}{2}\right)(2y-1)r^X r^Y(1-q_1)q_2\\&+(2x-1)\operatorname{sgn}\left(y-\frac{1}{2}\right)r^X r^Y q_1(1-q_2)\\&+\operatorname{sgn}\left(x-\frac{1}{2}\right)\operatorname{sgn}\left(y-\frac{1}{2}\right)r^X r^Y(1-q_1)(1-q_2).\end{aligned}\tag{4.11}$$

Since $\mathbb{E}(\varphi_1(X,Y,R^X,R^Y;m_X,m_Y))=0$, we can see that $\mathbb{V}\mathrm{ar}(\varphi_1(X,Y,R^X,R^Y;m_X,m_Y))$ is equal to $\mathbb{E}(\varphi_1(X,Y,R^X,R^Y;m_X,m_Y)^2)$. By squaring the sum in (4.11), and using mutual independence of R^X and R^Y, and their independence from X and Y, after integration, one can obtain

$$\mathbb{V}\mathrm{ar}(\varphi_1(X,Y,R^X,R^Y;m_X,m_Y))=\frac{1}{9}q_1q_2(3-3q_1+q_1^2)(3-3q_2+q_2^2).$$

By a known property of a non-degenerate U-statistic, we obtain that

$$\sqrt{n}\tilde{\tau}_n\xrightarrow{D}\mathcal{N}\left(0,\frac{4}{9}q_1q_2(3-3q_1+q_1^2)(3-3q_2+q_2^2)\right),$$

which completes the first step of the proof.

Second step. As seen in the Subsection 1.5.1, Randles (1982) obtained conditions under which the distribution of U-statistics with estimated parameters can be related to the distribution with the true parameter values. Due to having a very specific case in terms of notation, we restate the sufficient conditions 2.3 and 2.9A from the original paper.

(C1) [Orig. *2.3*] There is a neighborhood of (m_X,m_Y), called $K((m_X,m_Y))$ and a constant $K_1>0$ such that if $(m'_X,m'_Y)\in K((m_X,m_Y))$ and $D((m'_X,m'_Y),d)=(m'_X-d,m'_X+d)\times(m'_Y-d,m'_Y+d)$ is a rectangle centered at (m'_X,m'_Y) satisfying $D((m'_X,m'_Y),d)\subseteq K((m_X,m_Y))$, then

$$\begin{aligned}\mathbb{E}\Bigg(\sup_{(m''_X,m''_Y)\in D((m'_X,m'_Y),d)}\Big|&\Phi\big((X_1,Y_1,R_1^X,R_1^Y),(X_2,Y_2,R_2^X,R_2^Y);(m''_X,m''_Y)\big)\\&-\Phi\big((X_1,Y_1,R_1^X,R_1^Y),(X_2,Y_2,R_2^X,R_2^Y);(m'_X,m'_Y)\big)\Big|\Bigg)\leqslant K_1 d\end{aligned}\tag{4.12}$$

and

$$\begin{aligned}\lim_{d\to 0}\mathbb{E}\Bigg(\sup_{(m''_X,m''_Y)\in D((m'_X,m'_Y),d)}\Big|&\Phi\big((X_1,Y_1,R_1^X,R_1^Y),(X_2,Y_2,R_2^X,R_2^Y);(m''_X,m''_Y)\big)\\&-\Phi\big((X_1,Y_1,R_1^X,R_1^Y),(X_2,Y_2,R_2^X,R_2^Y);(m'_X,m'_Y)\big)\Big|^2\Bigg)=0.\end{aligned}\tag{4.13}$$

In the original Randles' formulation, $D((m'_X,m'_Y),d)$ is a disk, but our modification is equivalent due to the equivalence of norms in finite-dimensional space.

(C2) [Orig. *2.9A*] It holds that $\sqrt{n}((M_X, M_Y)-(m_X, m_Y)) = o_{\mathbb{P}}(1)$ and the expected value of the kernel $\Phi\big((X_i, Y_i, R_i^X, R_i^Y),(X_j, Y_j, R_j^X, R_j^Y);(m_X, m_Y)\big)$ has zero differential (with respect to (m_x, m_y)) at true parameter values.

Having these conditions satisfied, Randles (1982) proved that

$$\sqrt{n}\Big(\tilde{\tilde{\tau}}_n - \mathbb{E}\big(\Phi(\cdot;(\gamma_1,\gamma_2))\big)\Big|_{(\gamma_1,\gamma_2)=(M_X,M_Y)} - \tilde{\tau}_n + 0\Big) \xrightarrow{P} 0.$$

Expectations of terms in (4.10) go in pairs with opposite signs, so it is not difficult to note that $\mathbb{E}\Big(\Phi\big((X_i, Y_i, R_i^X, R_i^Y),(X_j, Y_j, R_j^X, R_j^Y);(M_X, M_Y)\big)\Big) = 0$. That means, if we show that conditions (C1) and (C2) are satisfied, we have obtained that the difference between $\sqrt{n}\tilde{\tilde{\tau}}_n$ and $\sqrt{n}\tilde{\tau}_n$ tends to zero in probability, so they have the same asymptotic distribution.

For the first condition, one can see that

$$\begin{aligned}
&\mathbb{E}\Bigg(\sup_{(m_X'',m_Y'')\in D((m_X',m_Y'),d)}\Big|\Phi\big((X_1, Y_1, R_1^X, R_1^Y),(X_2, Y_2, R_2^X, R_2^Y);(m_X'', m_Y'')\big)\\
&\qquad\qquad -\Phi\big((X_1, Y_1, R_1^X, R_1^Y),(X_2, Y_2, R_2^X, R_2^Y);(m_X', m_Y')\big)\Big|\Bigg)\\
&\leqslant \mathbb{E}\Big(\sup_{(m_X'',m_Y'')\in D((m_X',m_Y'),d)}\big|\operatorname{sgn}(X_1-m_X'')-\operatorname{sgn}(X_1-m_X')\big|\Big)\\
&+\mathbb{E}\Big(\sup_{(m_X'',m_Y'')\in D((m_X',m_Y'),d)}\big|\operatorname{sgn}(m_X''-X_2)-\operatorname{sgn}(m_X'-X_2)\big|\Big)\\
&+\mathbb{E}\Big(\sup_{(m_X'',m_Y'')\in D((m_X',m_Y'),d)}\big|\operatorname{sgn}(Y_1-m_Y'')-\operatorname{sgn}(Y_1-m_Y')\big|\Big)\\
&+\mathbb{E}\Big(\sup_{(m_X'',m_Y'')\in D((m_X',m_Y'),d)}\big|\operatorname{sgn}(m_Y''-Y_2)-\operatorname{sgn}(m_Y'-Y_2)\big|\Big)\\
&+\mathbb{E}\Big(\sup_{(m_X'',m_Y'')\in D((m_X',m_Y'),d)}\big|\operatorname{sgn}(X_1-m_X'')\operatorname{sgn}(Y_1-m_Y'')-\operatorname{sgn}(X_1-m_X')\operatorname{sgn}(Y_1-m_Y')\big|\Big)\\
&+\mathbb{E}\Big(\sup_{(m_X'',m_Y'')\in D((m_X',m_Y'),d)}\big|\operatorname{sgn}(m_X''-X_2)\operatorname{sgn}(Y_1-m_Y'')-\operatorname{sgn}(m_X'-X_2)\operatorname{sgn}(Y_1-m_Y')\big|\Big)\\
&+\mathbb{E}\Big(\sup_{(m_X'',m_Y'')\in D((m_X',m_Y'),d)}\big|\operatorname{sgn}(X_1-m_X'')\operatorname{sgn}(m_Y''-Y_2)-\operatorname{sgn}(X_1-m_X')\operatorname{sgn}(m_Y'-Y_2)\big|\Big)\\
&+\mathbb{E}\Big(\sup_{(m_X'',m_Y'')\in D((m_X',m_Y'),d)}\big|\operatorname{sgn}(m_X''-X_2)\operatorname{sgn}(m_Y''-Y_2)-\operatorname{sgn}(m_X'-X_2)\operatorname{sgn}(m_Y'-Y_2)\big|\Big). \qquad (4.14)
\end{aligned}$$

Begin by concentrating on the first term:

$$\begin{aligned}
&\mathbb{E}\Big(\sup_{(m_X'',m_Y'')\in D((m_X',m_Y'),d)}\big|\operatorname{sgn}(X_1-m_X'')-\operatorname{sgn}(X_1-m_X')\big|\Big)\\
&\qquad = \mathbb{E}\Big(\sup_{m_X''\in(m_X'-d,m_X'+d)}\big|\operatorname{sgn}(X_1-m_X'')-\operatorname{sgn}(X_1-m_X')\big|\Big).
\end{aligned}$$

Let us denote

$$\begin{aligned}
\xi_1(\omega) &= \sup_{m_X''\in(m_X'-d,m_X'+d)}\big|\operatorname{sgn}(X_1(\omega)-m_X'')-\operatorname{sgn}(X_1(\omega)-m_X')\big|\\
&= \sup_{m_X''\in(m_X'-d,m_X'+d)}\big|\operatorname{sgn}(X_1(\omega)-m_X'+(m_X'-m_X''))-\operatorname{sgn}(X_1(\omega)-m_X')\big|.
\end{aligned}$$

The supremum above can be 0 or 2. It will be equal to 2 if d is large enough so that adding (negative, if necessary) $m'_X - m''_X$ can change the sign of $X_1(\omega) - m'_X$, which happens when $|X_1(\omega) - m'_X| < d$. That gives us that

$$\mathbb{P}\{\xi_1 = 2\} = \mathbb{P}\{|X_1 - m'_X| < d\} = \mathbb{P}\{m'_X - d < X_1 < m'_X + d\} = F_X(m'_X + d) - F_X(m'_X - d).$$

Theorem 4.2 assumes that $f_X = F'_X$ is bounded in some neighborhood of the median. The set $K((m_X, m_Y))$ can be reduced if needed so that for any permissible m'_X and d it holds that $(m'_X - d, m'_X + d) \subseteq K((m_X, m_Y))$ lies in the mentioned neighborhood. By denoting $\frac{1}{2}K_1^{(1)}$ an upper bound for f_X, using Lagrange's mean value theorem we obtain

$$\mathbb{P}\{\xi_1 = 2\} = F_X(m'_X + d) - F_X(m'_X - d) \leqslant \frac{1}{2}K_1^{(1)} d.$$

Then, we get that

$$\mathbb{E}(\xi_1) = 2\mathbb{P}\{\xi_1 = 2\} \leqslant K_1 d,$$

which proves the condition (4.12) for the first term in (4.14). It is obvious that the same argument applies identically on all of the rest terms, producing constants $K_1^{(2)}, \ldots, K_1^{(8)}$. If we denote by $\xi_2, \ldots, \xi_n$ the rest of the suprema appearing in (4.14), we have that $\mathbb{E}(\xi_2) \leqslant K_1^{(2)} d, \ldots, \mathbb{E}(\xi_8) \leqslant K_1^{(8)} d$. Taking $K_1 = \max\{K_1^{(1)}, \ldots, K_1^{(8)}\}$, we have successfully proven (4.12). Condition (4.13) is verified in the exact same manner.

It is a known fact that the sample median has an asymptotic normal distribution. In a similar manner as Theorem 4.1, one can prove that $\sqrt{n}(M_X - m_X)$ has an asymptotic normal distribution, which means it is bounded in probability. The same argument can be used for M_Y, which is, by the assumed independence of the samples, enough to conclude that the first part of condition (C2) holds. Verification of the second part is also not a very difficult task, since the expected value of the kernel $\Phi\big((X_i, Y_i, R_i^X, R_i^Y), (X_j, Y_j, R_j^X, R_j^Y); (m_X, m_Y)\big)$ is itself zero, for any values m_X and m_Y, not just the true medians. Therefore, it has zero differential. This concludes the proof of Theorem 4.2. ■

The theoretical values q_1 and q_2 are unknown even to the imputer, and they are estimated by $\hat{q}_1 = \bar{R}_n^X = \frac{1}{n}\sum_{i=1}^n R_i^X$ and $\hat{q}_2 = \bar{R}_n^Y = \frac{1}{n}\sum_{i=1}^n R_i^Y$.

COROLLARY 4.2.1. *Denote $\hat{q}_1 = \bar{R}_n^X$ and $\hat{q}_2 = \bar{R}_n^Y$. Under assumptions of Theorem 4.2, it holds that, as $n \to \infty$,*

$$\frac{1}{\sqrt{\frac{4}{9}\hat{q}_1\hat{q}_2(3 - 3\hat{q}_1 + \hat{q}_1^2)(3 - 3\hat{q}_2 + \hat{q}_2^2)}} \sqrt{n}\tilde{\tilde{\tau}}_n \xrightarrow{D} \mathcal{N}(0, 1).$$

PROOF. Follows from Theorem 4.2, using that $\hat{q}_1$ and $\hat{q}_2$ are consistent estimators of q_1 and q_2, respectively, and then from Continuous Mapping Theorem and Slutsky's theorem. ■

REMARK 4.1. Similar results can be obtained when the mean imputation is used for handling missing data instead of median imputation. However, we note that the null distribution may no longer be free of the marginal distributions of X and Y.

4.3 Empirical study

In this section, we conduct an empirical study with the goal to compare $\hat{\tilde{\tau}}_n$ and $\tilde{\tilde{\tau}}_n$ as test statistics for testing independence, and to further explore the usability of obtained limiting results.

Bearing this in mind, we form critical regions relying on asymptotic normal distribution, i.e. we reject the null hypothesis when

$$\left|\frac{3}{2}\sqrt{\Sigma R_i}\hat{\tilde{\tau}}_n\right| \geqslant z_{\alpha/2} \quad \text{and} \quad \left|\frac{3}{2}\frac{\sqrt{n}\tilde{\tilde{\tau}}_n}{\sqrt{\hat{q}_1\hat{q}_2(3-3\hat{q}_1+\hat{q}_1^2)(3-3\hat{q}_2+\hat{q}_2^2)}}\right| \geqslant z_{\alpha/2}$$

where $z_{\alpha/2}$ is the upper quantile of the standard normal distribution at level α. We obtain empirical type I errors and powers of both $\hat{\tilde{\tau}}$ and $\tilde{\tilde{\tau}}$, for the level of significance $\alpha = 0.05$, and the rate of incomplete rows $1-q$, ranging from 0 to 30%, using the Monte Carlo approach with $N = 10000$ replicates. Here we consider the case of balance missing design, i.e. we assume that $q_1 = q_2$ (see the definition just before Theorem 4.2), while results for non-balanced designs with different values of q_1 and q_2 can be found in the Supplementary Material of Aleksić et al. (2023).

For ease of comparison, we also include the results for the complete sample case ($\hat{\tau}$). Empirical type I errors are calculated for sample sizes $n = 30, 50, 100$ and 200 when the marginal distributions are exponential $\mathscr{E}(1)$ (see Figure 4.1). The results for different marginals $\mathscr{E}(1)$ and $\mathscr{E}(2)$ can be seen from the aforementioned Supplementary Material.

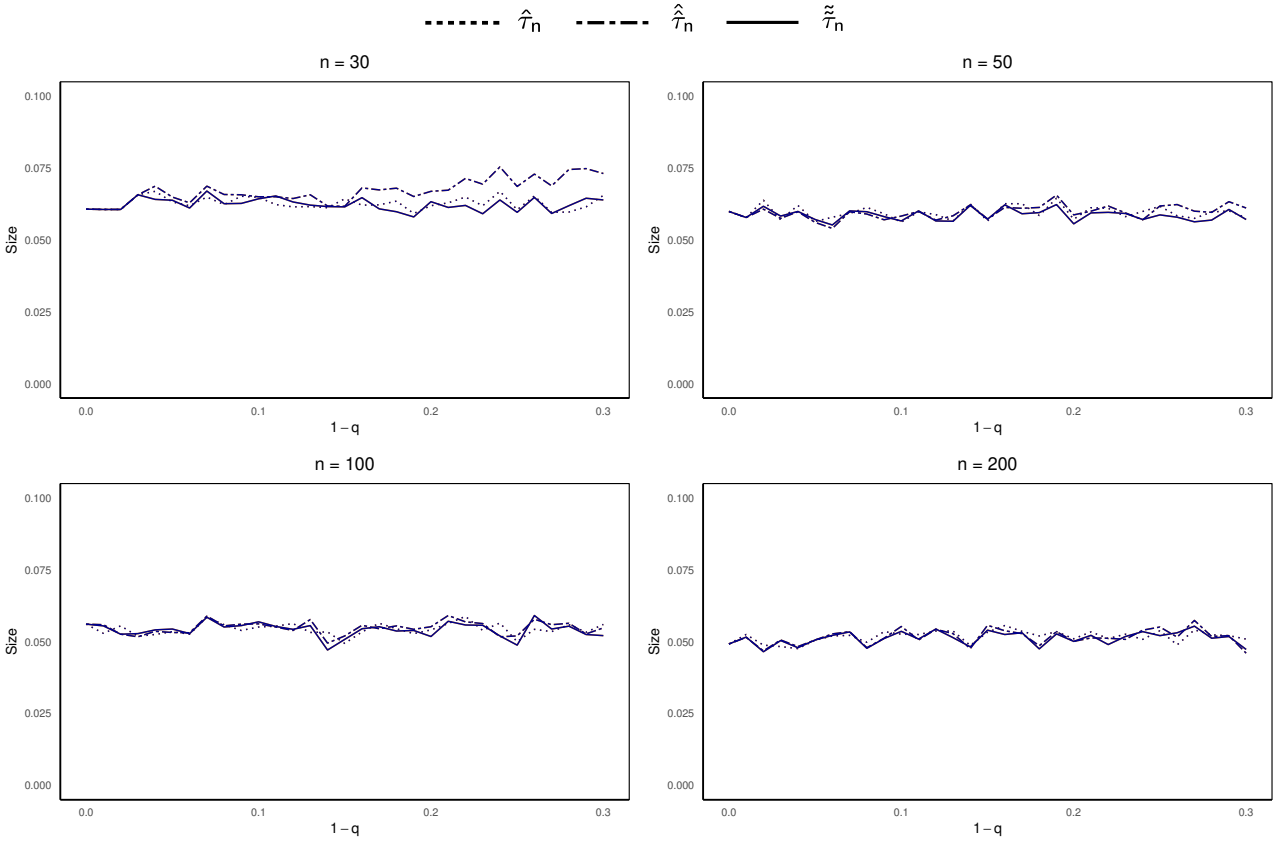

Figure 4.1: Empirical type I errors using two independent $\mathscr{E}(1)$ marginals, $q_1 = q_2$.

Figure 4.1 shows that, for samples of small size, even $\hat{\tau}_n$ exceeds the desired type I error of 0.05 with an empirical type I error of about 0.06. It is notable that $\tilde{\tilde{\tau}}_n$ exhibits similar behavior to $\hat{\tau}$ and does not experience a sudden increase in the type I error. On the other hand, it does not hold for $\hat{\tilde{\tau}}_n$. As the sample size increases, this problem becomes less notable for the sample size of $n = 50$ and $n = 100$. For $n = 200$ all three statistics are well calibrated. We note that in the case of non-balanced incompleteness design $\hat{\tilde{\tau}}_n$ and $\tilde{\tilde{\tau}}_n$ tests have the same behavior for all considered sample sizes.

From this point onward, we will compare $\hat{\tilde{\tau}}_n$ and $\tilde{\tilde{\tau}}_n$ in terms of power, for $n = 50$ and $n = 100$ and a wide range of alternatives, following the same study design as Cuparić and Milošević

(2024). We have not included the case of $n = 30$ since the tests under comparison are not of the same size. We note that there are ways to compare tests with different empirical sizes in terms of power, but such discussion falls out of the scope of our research. For a useful reference on that topic, see Batsidis et al. (2016). To ensure a point of reference, we will also include empirical powers obtained from $\hat{\tau}_n$.

In Figure 4.2 we present the power of the tests against the following bivariate distributions with $\mathscr{E}(1)$ marginals:

- Clayton copula, i.e. with survival function

$$\overline{H}(x,y) = 1 - \left(F_X(x)^{-\theta} + F_Y(y)^{-\theta} - 1\right)^{-\frac{1}{2}},\ \theta \in (0,\infty);$$

- Farlie-Gumbel-Morgenstern copula (labelled by FGM), i.e., with survival function

$$\overline{H}(x,y) = \overline{F}_X(x)\overline{F}_Y(y)(1+\theta F_X(x)F_Y(y)),\ \theta \in [-1,1].$$

Sample sizes of 50 and 100 are considered. Parameter θ is chosen in such a way that the Kendall's τ of the considered alternatives are equal to 0.1 or 0.2. From Figure 4.2 we can see that the powers decrease with the increase of missingness rate which is reasonable since the effective sample size decreases. In addition, the choice of the copula function has an impact on the behavior of the power. In the case of Clayton copula, the differences between complete-case and median-based imputation approaches are noticeable (but not drastically different), which does not hold for the FGM copula. In addition, the powers of tests, for fixed τ, and different choices of copula function slightly differ. That is even more evident from the figures presented in the Supplementary Material. From the figures presented therein, one can also see that the decrease in the power is the steepest in the balanced design.

4.4 Real-data example

To illustrate the proposed methodology, we consider a dataset from Kaggle: the Sales and satisfaction dataset (Mahmoudi, 2024). The dataset includes sales and customer satisfaction data from both before and after a specific intervention, along with purchase information for control and treatment groups. It contains 10000 observations on 7 variables: `Group` (14% missing values), indicating whether an observation belongs to the control or treatment group; `Customer_Segment` (20% missing values), that categorizes customers based on their value, as high, medium or low; `Sales_Before` (15% missing values), representing sales figures before the intervention; `Sales_After` (8% missing values), representing sales figures after the intervention; `Customer_Satisfaction_Before` (17% missing values), the customer satisfaction scores before the intervention on scale from 1 to 100; `Customer_Satisfaction_After` (16% missing values), the customer satisfaction scores after the intervention; `Purchase_Made` (8% missing values), variable indicating whether a customer made a purchase.

It is natural and interesting to see whether there exists any relation between the customer satisfaction and sales figures. For that purpose, we considered two pairs of variables: `Customer_Satisfaction_Before` and `Sales_Before`, as well as `Customer_Satisfaction_After` and `Sales_After`. To apply our proposed methodology, as well as complete-case, on those variables, we first need to verify that the MCAR assumption holds. In addition to our generalized MCAR test developed in Section 3.3 (applied to centered data), we employ the compatibility-based test by Bordino and Berrett (2024) and the one by Little (1988). We note, however, that Little's test relies on the assumption of normality and is highly sensitive to departures from this assumption, as demonstrated in Chapter 3. This will be the reason not to use Little's test in

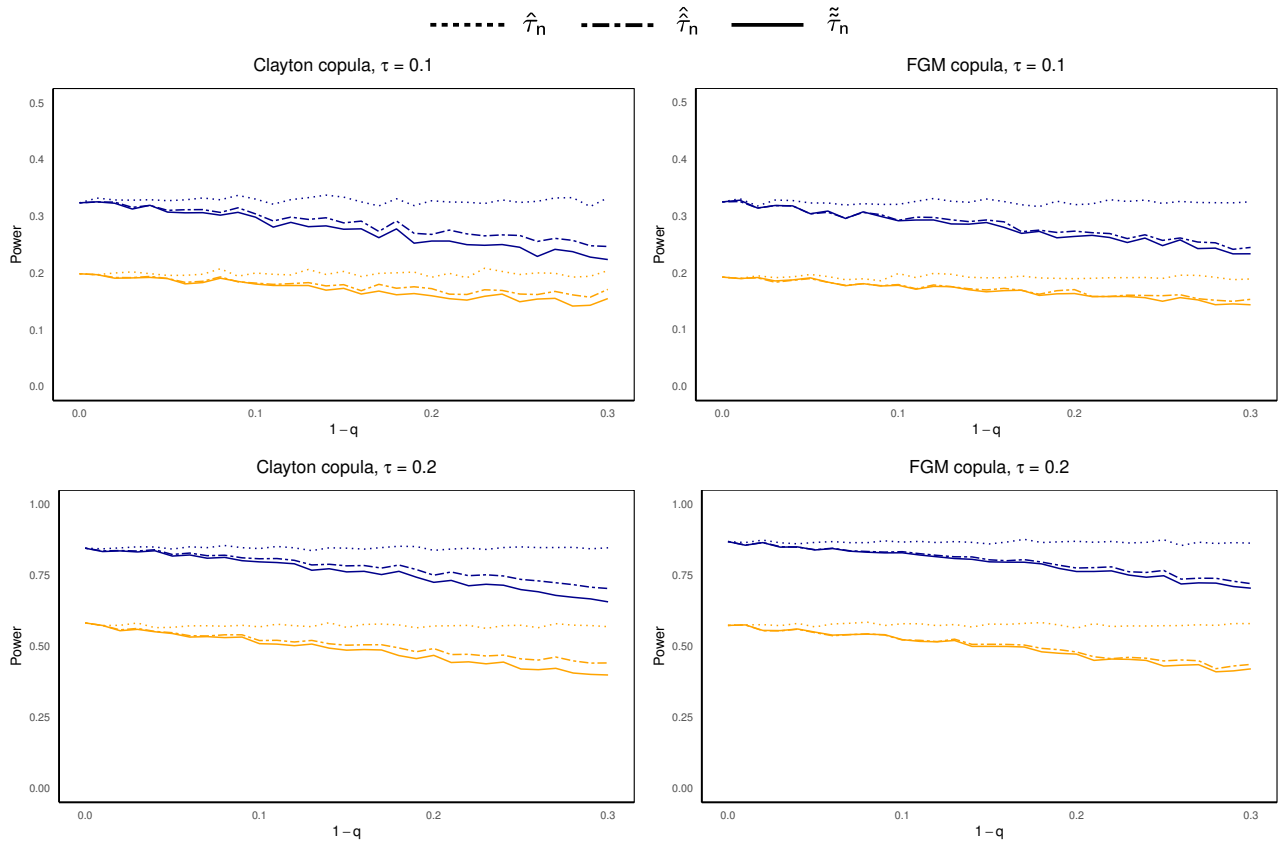

Figure 4.2: Empirical test powers using Clayton and FGM copulas with $\mathcal{E}(1)$ marginals for $n = 50$ (bottom group, in orange) and $n = 100$ (top group, in blue), $q_1 = q_2$.

the following chapter, where we test multivariate normality on an incomplete dataset. The results of MCAR testing, both on the complete dataset as well as the two variables of interest, are presented in the Table 4.1. As we can see, there is no reason to reject the MCAR assumption.

Table 4.1: Sales and satisfaction dataset: p-values of MCAR tests for the entire dataset, pair `Customer_Satisfaction_Before` and `Sales_Before`, and pair `Customer_Satisfaction_After` and `Sales_After`.

Variables used / Test	Aleksić (Section 3.3)	Bordino and Berrett (2024)	Little (1988)
All variables	0.984	1.000	0.345
Pair before	0.838	1.000	0.838
Pair after	0.532	1.000	0.532

Proceeding with the independence testing, both complete-case and median imputation approach detect strong relationship between variables and reject the null hypothesis of independence with p-value effectively equal to zero.

To see the behavior of the methods on real-world dataset of smaller size, we randomly select a subsample of size 100 from the dataset. Although we have seen from Table 4.1 that there is no evidence against MCAR from the entire dataset, for methodological purposes we redo the MCAR testing on the selected subsample. The results can be seen in Table 4.2. Again, we do not reject the MCAR assumption.

When we conduct the Kendall's test on the pair, we obtain the p-values seen in Table 4.3. As we can see, both studied approaches do not reject the null hypothesis of independence (i.e. $\tau = 0$) of variables `Customer_Satisfaction_Before` and `Sales_Before`.

However, for variables `Customer_Satisfaction_After` and `Sales_After`, we have different conclusions. At the standard significance level of 0.05, complete-case approach suggests

Table 4.2: Sales and satisfaction dataset (subsample of size 100): p-values of MCAR tests for the entire dataset, pair `Customer_Satisfaction_Before` and `Sales_Before`, and pair `Customer_Satisfaction_After` and `Sales_After`.

Variables used / Test	Aleksić (Section 3.3)	Bordino and Berrett (2024)	Little (1988)
All variables	0.719	1.000	0.332
Pair before	0.934	0.949	0.611
Pair after	0.977	1.000	0.944

Table 4.3: Sales and satisfaction dataset (subsample of size 100): p-values of Kendall's test under the complete-case and median imputation approach for handling missing data.

Pair of variables/Approach for handling missing data	Complete-case	Median imputation
Before	0.1615	0.159
After	0.021	0.105

rejecting the null hypothesis, and median imputation approach suggests the opposite. Having the results of our simulation study, we know that for sample size of 100 both approaches should be well calibrated and the complete-case should be slightly more powerful. In this case, we would reject the null hypothesis and state that there is some correlation between the Customer satisfaction and the Sales numbers after the intervention.

Chapter 5

The BHEP test for MCAR data

Although the problem of missing data has been studied from various perspectives, in the context of goodness-of-fit testing, the literature is very sparse. On the other hand, testing multivariate normality is a crucial aspect of statistics as it provides a foundation for many statistical techniques and assumptions. In multivariate data analysis, researchers often assume that the data follows a multivariate normal distribution. Deviations from multivariate normality (MVN) can affect the validity of various statistical methods, such as multivariate analysis of variance, some properties of coefficients of linear regression, etc. Identifying departures from MVN allows researchers to make informed decisions about the appropriateness of chosen statistical methods and can guide the selection of alternative techniques if necessary. Ensuring MVN is also important in fields like finance, biology, and social sciences, where accurate modeling of data distributions is essential for having valid inference and quality decisions based on statistical analyses. For this reason, many tests for testing MVN have been proposed so far. Mecklin and Mundfrom (2005) gave a useful comparison in terms of empirical type I and type II error rates, while for some more recent results we refer to Ebner and Henze (2020), Ebner et al. (2022), González-Estrada et al. (2022), and Ejsmont et al. (2023).

Only a handful of tests were proposed for testing MVN on an incomplete sample, and they are based on skewness and kurtosis estimates. Yamada et al. (2015) proposed a test for MVN that is based on a generalization of Mardia's statistic for measuring kurtosis. The test was suitable for two-step monotone incomplete data, which is a special case of missing data. Another such test was given by Kurita and Seo (2022). Tan et al. (2005) proposed a test for multiple samples of possibly small sizes, that was, again, based on estimating kurtosis and skewness, utilizing multiple imputation and Gibbs sampler. Recently, an extended simulation study was conducted by Tsatsi et al. (2024) to compare various tests (including BHEP) under numerous imputation methods, for two-step monotone MCAR data. To the best of our knowledge, MVN testing for arbitrary patterns of MCAR data has not been studied in the context of the BHEP test, nor within the broader class of tests with L^2-weighted test statistics, to which the BHEP test belongs. We aim to fill that gap with results from this chapter, which were obtained by Aleksić and Milošević (2025a).

The remainder of this chapter is structured as follows.

In Section 5.1 we make a brief review of the BHEP test and its properties that are essential for studying the test in the presence of missing data, which is one of our aims. Sections 5.2, 5.3, and 5.4 are structured to highlight the key contributions of this chapter.

- In Subsection 5.2.1 we examine the behavior of the BHEP test statistics under the complete-case approach, and we prove that it is suitable for testing MVN under the MCAR assumption. Specifically, we show that the asymptotic distribution of the BHEP test statistic based on complete cases is the same as that based on the full sample.

- The contribution of Subsection 5.2.2 is to give an insight in the technique that might

be used for deriving the asymptotic distribution of the BHEP test statistic calculated on an imputed dataset. We explain the complexity of the derivations, even for the simplest imputation techniques.

- Section 5.3 offers an alternative approach, by presenting a bootstrap algorithm that approximates the null distribution of the test statistic calculated on an imputed sample. Then, it presents the results of an extensive simulation study that compares the type I error preservation and power behavior of the test statistic under the complete-case approach, as well as with some of the most widely used imputation methods, using the proposed bootstrap algorithm. A discussion of the findings is also provided.
- Section 5.4 clearly illustrates all of the discussed approaches using a real-data example. Additionally, the section discusses the p-values obtained from the real data, taking into account the results of the simulations from the previous section.

5.1 Prerequisites

Here, we focus on the BHEP test which is one of the most well-known procedures for testing MVN assumption and review its properties in the absence of missing data. For more details, we refer to Baringhaus and Henze (1988) and to Henze and Wagner (1997) for the generalization of the test. All of the results stated here are known and will be utilized afterwards.

Let $X_1, X_2, \ldots, X_n$ be a sample of n IID d-variate random column vectors, equally distributed as $X = \left(X^{(1)}, \ldots, X^{(d)}\right)^T$. We are interested in testing the assumption that X has some d-variate non-singular normal distribution, i.e. the hypothesis

$$H_d : \text{the law of } X \text{ is } \mathcal{N}_d(\mu, \Sigma),$$

for some $\mu \in \mathbb{R}^d$ and some non-singular covariance matrix Σ, where $\mathcal{N}_d(\mu, \Sigma)$ denotes the d-variate normal distribution with mean μ and covariance matrix Σ.

Let

$$S_n = \frac{1}{n} \sum_{j=1}^{n} (X_j - \bar{X}_n)(X_j - \bar{X}_n)^T$$

be the sample covariance matrix, and let

$$Y_j = S_n^{-\frac{1}{2}} (X_j - \bar{X}_n), \quad j = 1, 2, \ldots, n.$$

If we denote by

$$\psi_n(t) = \frac{1}{n} \sum_{j=1}^{n} \exp\left(i t^T Y_j\right), \quad t \in \mathbb{R}^d$$

the empirical characteristic function of $Y_1, \ldots, Y_n$, the test statistic of the BHEP test is given as

$$T_n = \int_{\mathbb{R}^d} \left| \psi_n(t) - \exp\left(-\frac{1}{2}\|t\|^2\right) \right|^2 \varphi(t)\, \mathrm{d}t, \tag{5.1}$$

where $\varphi(t) = (2\pi)^{-d/2} \exp(-\frac{1}{2}|t|^2)$.

Baringhaus and Henze (1988) were successful in proving that T_n can be represented as a weakly degenerate V-statistic with estimated parameters. More precisely, they have shown that

$$T_n = V_n(\lambda_n),$$

where $\lambda = (\mu, \Sigma)$, $\lambda_n = (\bar{X}_n, S_n)$, $V_n(\lambda) = n^{-2}\sum_{j,k=1}^{n} h(X_j, X_k; \lambda)$ and the closed-form expression for the kernel h is given as

$$\begin{aligned} h(x_1, x_2; \lambda) = &\exp\Big(-\frac{1}{2}(x_1 - x_2)^T\Sigma^{-1}(x_1 - x_2)\Big) \\ &-2^{-d/2}\exp\Big(-\frac{1}{4}(x_1 - \mu)^T\Sigma^{-1}(x_1 - \mu)\Big) \\ &-2^{-d/2}\exp\Big(-\frac{1}{4}(x_2 - \mu)^T\Sigma^{-1}(x_2 - \mu)\Big) + 3^{-d/2}, \end{aligned} \tag{5.2}$$

or, as they have shown

$$h(x_1, x_2; \lambda) = \int_{\mathbb{R}^d} g(x_1, t; \lambda) g(x_2, t; \lambda)\varphi(t)\,\mathrm{d}t, \tag{5.3}$$

where

$$g(x, t; \lambda) = \cos\big(t^T\Sigma^{-1/2}(x - \mu)\big) + \sin\big(t^T\Sigma^{-1/2}(x - \mu)\big) - \exp\Big(-\frac{1}{2}\|t\|^2\Big).$$

The authors then used existing theory about weakly degenerate U- and V-statistics with estimated parameters that was developed by De Wet and Randles (1987) to derive the asymptotic distribution of T_n and to prove some further properties of the test. Specifically, the authors found that

$$n T_n \xrightarrow{D} \sum_{k=1}^{+\infty} \kappa_k \chi^2_{1,k}, \tag{5.4}$$

where $\chi^2_{1,k}$ are IID χ^2_1-distributed random variables, and κ_k-s are eigenvalues of the integral operator A defined as

$$Af(x) = \int_{\mathbb{R}^d} h_*(x, y) f(y)\varphi(y)\,\mathrm{d}y. \tag{5.5}$$

The closed form of the kernel h_* was derived by Baringhaus and Henze (1988), while the problem of deriving the eigenvalues was recently studied by Ebner and Henze (2023).

Although the asymptotic distribution (5.4) can be approximated, in practice, the null distribution is usually approximated using Monte Carlo simulation. Since the T_n is affine invariant, one can safely assume that the null distribution is d-variate standard normal, and simulate the null distribution for the desired sample size. The empirical quantiles are then used to construct the rejection region of the test.

5.2 Challenges of incomplete datasets

Now, assume that we have a sample $X_1, \ldots, X_n$ of IID d-variate random column vectors and that some of the data are missing. For every $j = 1, 2, \ldots, n$ and $k = 1, 2, \ldots, d$, the *response indicator* for $X_j^{(k)}$ will be defined here as:

$$R_j^{(k)} = \begin{cases} 1, & \text{if } X_j^{(k)} \text{ is observed,} \\ 0, & \text{otherwise,} \end{cases}$$

where $X_j^{(k)}$ is k-th component of X_j, as in previous chapters. Denote $R_j = (R_j^{(1)}, \ldots, R_j^{(d)})^T$. We assume that R_js are mutually independent, and denote $q = (q_1, \ldots, q_d)^T = \mathbb{E}(R_1)$. Finally, we use $\odot$ to denote the Hadamard–Schur componentwise multiplication of vectors.

In what follows we consider the expanded sample

$$(X_1^T, R_1^T)^T, \ldots, (X_n^T, R_n^T)^T,$$

which is suitable for expressing the test statistic in the presence of missing data.

5.2.1 Complete-case approach

The complete-case approach, i.e. when every observation that is not completely observed is removed from the sample, is very common in practice. It is mostly used when data are MCAR and the missingness rate is not very high, since most estimates, such as sample mean, remain unbiased and consistent. Intuitively, one can expect that a test statistic will preserve properties under MCAR and complete-case, since we are working with the "representative" subset of the original sample. The claim remains true for the test statistic of the BHEP test. Although intuitive, it deserves rigorous proof. Furthermore, the proof of a similar result for non-degenerate U-statistics was anything but trivial, as shown in Chapter 4.

The proof relies on the results given by De Wet and Randles (1987), which we restated in Subsection 1.5.2, slightly adjusted to our notation.

THEOREM 5.1 [ALEKSIĆ, MILOŠEVIĆ (2024)]**.** *Let the $X_1, \ldots, X_n$ be IID sample from non-degenerate d-variate normal distribution, let the MCAR assumption hold, and let T_n be as defined in* (5.1)*. Let $\hat{T}_n$ be the same statistic, but calculated on the completely observed sample units. Then, it holds that nT_n and $\hat{n}\hat{T}_n$ have the same asymptotic distribution, where $\hat{n}$ is the number of complete cases.*

PROOF. First, observe that $q = \mathbb{E}(R_1)$ is not treated as a distributional parameter, but rather as a known constant. The test statistic does not use its estimate in the sense of De Wet and Randles.

Secondly, note that $\hat{T}_n$ is also affine invariant with respect to transformations of $X_1, \ldots, X_n$. It follows in the same manner as for the T_n. For more detail, one can consult Baringhaus and Henze (1988).

Now, if we denote by $S_j = \prod_{k=1}^{d} R_j^{(k)}$ an indicator that X_j is completely observed, one can easily see that, if $\hat{n} = \sum_{j=1}^{n} S_j$, then $\hat{n}/n \xrightarrow{P} q_\Pi$ under MCAR, where $q_\Pi = \mathbb{E}(S_1)$. From this point onwards, we can consider that we work with the expanded sample

$$(X_1^T, S_1)^T, \ldots, (X_n^T, S_n)^T.$$

If we introduce

$$\hat{g}((x,s),t;\lambda) = s g(x,t;\lambda),$$

and

$$\hat{h}((x,s_x),(y,s_y);\lambda) = \int_{\mathbb{R}^d} \hat{g}((x,s_x),t;\lambda)\hat{g}((y,s_y),t;\lambda)\varphi(t)\,\mathrm{d}t = s_x s_y h(x,y;\lambda),$$

one can see that $\hat{\epsilon}(t;\lambda) := \mathbb{E}(\hat{g}((X,S),t;\lambda)) = q_\Pi \mathbb{E}(g(X,t;\lambda)) = q_\Pi \epsilon(t;\lambda)$, where $\epsilon(t;\lambda)$ is the same as from Baringhaus and Henze (1988). Having these direct relations between kernels makes regularity conditions of De Wet and Randles (1987) (see Subsection 1.5.2) follow trivially from those derived by Baringhaus and Henze (1988). The only one that requires attention is Condition 2.10, i.e. the expansion of parameter estimates, which we now present.

The parameters here are estimated on the complete cases, being

$$\hat{\mu} = \frac{1}{\hat{n}} \sum_{j=1}^{n} S_j X_j = \frac{1}{n} \sum_{j=1}^{n} \frac{S_j}{q_\Pi} X_j + o_{\mathbb{P}}(1/\sqrt{n})$$

and

$$\hat{\Sigma} = \frac{1}{\hat{n}} \sum_{j=1}^{n} S_j (X_j - \bar{X}_j)(X_j - \bar{X}_j)^T = \frac{1}{n} \sum_{j=1}^{n} \frac{S_j}{q_\Pi} (X_j - \bar{X}_j)(X_j - \bar{X}_j)^T + o_{\mathbb{P}}(1/\sqrt{n}).$$

This allows us to use

$$\hat{\alpha}(x, s) = \frac{s}{q_\Pi} \alpha(x)$$

to express

$$(\hat{\mu}, \hat{\Sigma}) = (0, I_d) + \frac{1}{n} \sum_{j=1}^{n} \hat{\alpha}(X_j, S_j) + o_{\mathbb{P}}(1/\sqrt{n}),$$

where $\alpha(x) = (x, x x' - I_d)$. Having this, and the direct relation $\hat{\epsilon}_1(t; 0, I_d) = q_\Pi \epsilon_1(t; 0, I_d)$, where ϵ_1 is the same as one from Baringhaus and Henze (1988), we proceed in the same manner and obtain the kernel

$$\begin{aligned} \hat{h}_*((x, s_x), (y, s_y)) = \int_{\mathbb{R}^d} & \left(\hat{g}((x, s_x), t; 0, I_d) - \hat{\epsilon}_1(t; 0, I_d) \hat{\alpha}(x, s_x) \right) \cdot \\ & \cdot \left(\hat{g}((y, s_y), t; 0, I_d) - \hat{\epsilon}_1(t; 0, I_d) \hat{\alpha}(y, s_y) \right) \varphi(t) \, \mathrm{d}t \\ = s_x s_y h_*(x, y). & \end{aligned} \tag{5.6}$$

By De Wet and Randles, and similarly to Baringhaus and Henze, we have that

$$n \hat{T}_n \xrightarrow{D} \sum_{k=1}^{+\infty} \zeta_k \chi^2_{1,k},$$

where ζ_k are the eigenvalues of an integral operator

$$\begin{aligned} B g(x, s_x) &= \sum_{s_y \in \{0,1\}} \left(\int_{\mathbb{R}^d} s_x s_y h_*(x, y) g(y, s_y) \varphi(y) \, \mathrm{d}y \right) \mathbb{P}\{S = s_y\} \\ &= s_x q_\Pi \int_{\mathbb{R}^d} h_*(x, y) g(y, 1) \varphi(y) \, \mathrm{d}y. \end{aligned}$$

Now we see that $g(x, s_x)$ is an eigenfunction of B if and only if $g(x, s_x) = s_x f(x)$, where $f(x)$ is an eigenfunction of A, where A is from (5.5). Furthermore, every eigenvalue of B is the eigenvalue of A multiplied by q_Π. Then, it follows that

$$n \hat{T}_n \xrightarrow{D} q_\Pi \sum_{k=1}^{+\infty} \kappa_k \chi^2_{1,k},$$

where the right-hand side is as in (5.4). Since $n/\hat{n} \xrightarrow{P} 1/q_\Pi$, Slutsky's theorem gives us that

$$\hat{n} \hat{T}_n \xrightarrow{D} \sum_{k=1}^{+\infty} \kappa_k \chi^2_{1,k},$$

which concludes the proof. ■

5.2.2 Imputation approach

In this subsection, we discuss the BHEP test in the context of the imputed dataset, illustrating the methodology and problems that arise using the example of sample mean imputation. To be more formal, we replace every $X_j^{(k)},\ j=1,2,\ldots,n,\ k=1,2,\ldots,d$, that is not observed, with the complete-case estimator of its mean $\frac{1}{\sum_{j=1}^n R_j^{(k)}}\sum_{j=1}^n X_j^{(k)}R_j^{(k)}$, calculated on available column units. Now, one might wish to derive the asymptotic distribution of the BHEP test statistic calculated on the imputed sample. To be able not to work from scratch, but to rely on some known results, it is wise to first use components of the theoretical mean value $\mu=(\mu_1,\ldots,\mu_d)^T$ of the distribution as the imputed values, and then move to estimated ones. This is due to the fact that we can see μ as the parameters of the kernel of the statistic, and then rely on the results of De Wet and Randles (1987) (see Subsection 1.5.2) to be able to replace them with sample means.

By noting that instead of saying that data value is getting replaced with corresponding μ_i when missing, and remaining unchanged when observed, one can neatly write that every X_j from the sample, $j=1,\ldots,n$, is being replaced with

$$(X_j-\mu)\odot R_j+\mu.$$

Now, the test statistic on the (sample) imputed dataset can be written as

$$\tilde{T}_n = n\,\tilde{V}_n(\tilde{\lambda}_n),$$

where $\tilde{\lambda}_n=(\tilde{X}_n,\tilde{S}_n)$ are parameters estimated from the incomplete data. The vector of means is estimated as mentioned above, i.e. $\tilde{X}_n=\frac{1}{\sum_{j=1}^n R_j}\sum_{j=1}^n X_j\odot R_j$, where we use slight abuse of notation since the division is also componentwise, R_j being a vector for every j. The covariance matrix Σ can be estimated using only complete cases:

$$\tilde{\Sigma}=\frac{1}{\sum_{j=1}^n S_j}\sum_{j=1}^n S_j(X_j-\bar{X}_n)(X_j-\bar{X}_n)^T. \tag{5.7}$$

Going further, one can see that the auxiliary V-statistic $\tilde{V}_n(\lambda)$ can be written as $\tilde{V}_n(\lambda)=n^{-2}\sum_{j,k=1}^n \tilde{h}(X_j,X_k;\lambda)$, where

$$\tilde{h}((x_1,r_1),(x_2,r_2);\lambda)=h\big((x_1-\mu)\odot r_1+\mu,(x_2-\mu)\odot r_2+\mu;\lambda\big) \tag{5.8}$$

and, similarly

$$\tilde{g}((x,r),t;\lambda)=g\big((x-\mu)\odot r+\mu,t;\lambda\big).$$

Since integrations depend only on t, the relation (5.3) holds for $\tilde{h}$ and $\tilde{g}$. To rely on the known results for the asymptotics of weakly degenerate V-statistics with estimated parameters, one needs to verify that conditions of De Wet and Randles (1987) hold.

It is readily seen that

$$\begin{aligned}\tilde{g}\big((x,r),t;\lambda\big)&=\cos\big(t^T\Sigma^{-1/2}((x-\mu)\odot r)\big)+\sin\big(t^T\Sigma^{-1/2}((x-\mu)\odot r)\big)-\exp\left(-\frac{1}{2}\|t\|^2\right)\\&=\cos\big(t^T\Sigma^{-1/2}(x\odot r)\big)\cos\big(t^T\Sigma^{-1/2}(\mu\odot r)\big)+\sin\big(t^T\Sigma^{-1/2}(x\odot r)\big)\sin\big(t^T\Sigma^{-1/2}(\mu\odot r)\big)\\&\quad+\sin\big(t^T\Sigma^{-1/2}(x\odot r)\big)\cos\big(t^T\Sigma^{-1/2}(\mu\odot r)\big)-\cos\big(t^T\Sigma^{-1/2}(x\odot r)\big)\sin\big(t^T\Sigma^{-1/2}(\mu\odot r)\big)\\&\quad-\exp\left(-\frac{1}{2}\|t\|^2\right).\end{aligned} \tag{5.9}$$

If we go back to the kernel h as in (5.2) and substitute every missing X_j with sample mean, and looking at $\tilde{\Sigma}$ as in (5.7), we can easily see that $\tilde{V}_n(\tilde{\lambda}_n)$ is invariant with respect to translations, i.e. does not depend on the expected value of the data. Unfortunately, it does depend on the covariance matrix of the data. The discussion can be found in the Supplementary Material of Aleksić and Milošević (2025a). From this point onward, it will be assumed that data are IID sampled from the d-variate normal distribution $\mathcal{N}_d(0,\Delta)$, for some positive definite covariance matrix Δ.

Remark 5.1. Note that for the column vectors a, b, c from $\mathbb{R}^d$ it holds that $a^T(b \odot c) = (a^T \odot c^T)b$.

Following the idea of Baringhaus and Henze (1988), let us focus on the first term of the first summand of (5.9), that is $\cos\big(t^T\Sigma^{-1/2}(x \odot r)\big)$ (the other term is constant with respect to x). Assuming for a moment that R is a constant vector equal to r, one is able, relying on the Remark 5.1, to calculate that

$$\begin{aligned}\mathbb{E}\Big(\cos\big(t^T\Sigma^{-1/2}(X \odot R)\big)\Big| R = r\Big) &= \int_{\mathbb{R}^d} \cos\big(t^T\Sigma^{-1/2}(x \odot r)\big) f_{\mathcal{N}_d(0,\Delta)}(x)\,\mathrm{d}x \\ &= \int_{\mathbb{R}^d} \cos\Big(\big((t^T\Sigma^{-1/2}) \odot r^T\big)x\Big) f_{\mathcal{N}_d(0,\Delta)}(x)\,\mathrm{d}x,\end{aligned}$$

which is exactly the characteristic function of the $\mathcal{N}_d(0,\Delta)$ distribution, calculated at the point $(t^T\Sigma^{-1/2}) \odot r^T$, so we obtain

$$\mathbb{E}\Big(\cos\big(t^T\Sigma^{-1/2}(X \odot R)\big)\Big| R = r\Big) = \exp\Big(-\frac{1}{2}\big((t^T\Sigma^{-1/2}) \odot r^T\big)\Delta\big((t^T\Sigma^{-1/2}) \odot r^T\big)^T\Big).$$

Conducting similar calculations on the other terms in (5.9), one can obtain that

$$\begin{aligned}&\mathbb{E}\Big(\tilde{g}\big((X,R),t;\lambda\big)\Big| R = r\Big) \\ &\qquad = \Big(\cos\big(t^T\Sigma^{-1/2}(\mu \odot r)\big) - \sin\big(t^T\Sigma^{-1/2}(\mu \odot r)\big)\Big) \\ &\qquad\quad \cdot \exp\Big(-\frac{1}{2}\big((t^T\Sigma^{-1/2}) \odot r^T\big)\Delta\big((t^T\Sigma^{-1/2}) \odot r^T\big)^T\Big) - \exp\Big(-\frac{1}{2}\|t\|^2\Big).\end{aligned}$$

The next step would be to obtain the expected value with respect to R of the above expression, which is a finite sum over all possible values of d-tuples of zeros and ones.

To be able to use conditions of De Wet and Randles (1987), as we did in proving Theorem 5.1, one needs to verify conditions 2.9–2.11 of De Wet and Randles (1987) (see Subsection 1.5.2). On the first glance, calculations go as smoothly as before, with the only difference being that we do not use $\hat{g}$ and its expected value $\hat{\epsilon}$, but $\tilde{g}$ and its expected value

$$\tilde{\epsilon}(t,\lambda) = \mathbb{E}\big(\tilde{g}\,((X,R),t;\lambda)\big).$$

Then, assuming the underlying $\mathcal{N}_d(0,\Delta)$ distribution, we need to find the vector of its partial derivatives, calculated at the true value of parameters (here $\lambda = (0,\Delta)$), denoted by $\tilde{\epsilon}_1(t;\lambda)$. This would then be used to find the kernel similar to (5.6), and to determine the asymptotic distribution using its eigenvalues.

However, further calculations would be of no great help, since this kernel, and subsequently, its operator eigenvalues, depend on the unknown distribution parameters, and it is not very likely that one could be able to obtain them analytically. In the complete-data case, the null distribution can be simulated by sampling from the multivariate standard normal distribution, and empirical quantiles can be used for determining the critical values. This is all

due to the fact that original test statistic proposed by Baringhaus and Henze (1988) is affine invariant, and consequently distribution-free. Our statistic is neither, so we believe that, at this point, it is more convenient to rely on resampling methods for testing MVN in this context.

Another (more natural) choice would be to estimate Σ as the sample covariance matrix calculated on the imputed dataset, but for that estimate, it is difficult to verify the Condition 2.10 of De Wet and Randles (1987) whose results one may aim to utilize in this context. However, our preliminary simulations indicate that if that choice is made, for imputation methods used in this study, the null distribution of test statistic does not depend on the location and scale parameters of the underlying multivariate normal distribution (see the Supplementary Material of Aleksić and Milošević (2025a)). Up to this point, proving so remains an open question.

5.3 Empirical study

In this section, an extensive simulation study is conducted to answer a crucial question that motivated our work: Is it better to impute the data, or to use a complete-case approach when using BHEP test of MVN? We observe different scenarios, varying underlying data distribution, missingness rate, as well as imputation methods. Here we highlight the most significant simulation results that reinforce our key points. Additional results can be found in the Supplementary Material of Aleksić and Milošević (2025a).

At this point, as seen before, we have only empirical indications that the distribution of the BHEP test statistic calculated on the imputed dataset does not depend on the mean and covariance parameters, and only for data scaled using parameter estimates obtained from the imputed data. Moreover, this distribution does depend on the missingness rate. Having that in mind, we offer the Algorithm 5.1, that simulates the distribution of the BHEP test statistic calculated on the imputed data, assuming the MCAR missingness, and is able to utilize various imputation approaches.

Remark 5.2. We note that the Algorithm 5.1 is a modification of a bootstrap algorithm proposed by Jiménez-Gamero et al. (2003), that was designed to work with a complete sample. One of the goals of the empirical study that follows is to examine its properties under various methods of imputation.

Empirical type I errors and powers are obtained using Monte Carlo procedure where Algorithm 5.1 is repeated $N = 2000$ times with $B = 1000$ bootstrap cycles in each repetition. The latter is a common value in literature (e.g. Jiménez-Gamero and Alba-Fernández, 2021). In order to make a fair comparison we apply the same procedure along different approaches, although we are aware that the usage of bootstrap in the complete-case is not necessary. Everything is done for the level of significance $\alpha = 0.05$.

The choice of imputation methods came down to mean imputation, median imputation, as well as 3- and 6-nearest neighbor imputation. For the first two, built-in functions from the `R` `missMethods` package are used (Rockel, 2023), and for the kNN we use `knn.impute` function from the package `bnstruct` (Franzin et al., 2017).

MCAR data are created using the `delete_MCAR` function from the `missMethods` package. To examine the properties of each method, with the increase in sample size, sample sizes of 30, 60, 90 and 120 are considered. As the alternatives to the null hypothesis, we consider Student's t distribution with several degrees of freedom. This family is frequently used as an alternative distribution within the literature related to testing MVN (Ebner and Henze, 2020; Ejsmont et al., 2023). In particular, we look for the power behavior against Student's t distribution with 5, 7 and 11 degrees of freedom. Also, 2-dimensional and 3-dimensional data are generated, and the covariance/scale matrix is also varied. Here, we present empirical type I errors for standard normal distribution, while the cases of other covariance matrices, with their descriptions, are

Algorithm 5.1 A bootstrap algorithm for testing MVN on an incomplete sample with MCAR data

1: Start with the incomplete sample $x = (x_1, \ldots, x_n)$;
2: Obtain the imputed dataset x_{imp} using the chosen method;
3: Obtain the value of test statistic $T_n(x_{\text{imp}})$ on the imputed dataset; that is, the data are standardized using the sample covariance matrix and the sample mean calculated on the imputed dataset;
4: Estimate covariance matrix Σ by $\tilde{\Sigma}$, calculated on the complete cases; estimate mean vector μ by $\tilde{\mu}$ on the dataset that is imputed by the chosen method;
5: Estimate $\tilde{p}$ of the vector of by-column missingness probabilities using response indicator averages;
6: Generate bootstrap sample $x^* = (x_1^*, \ldots, x_n^*)$ from $\mathcal{N}_d(\tilde{\mu}, \tilde{\Sigma})$;
7: Generate missingness in x^* according to MCAR and probabilities $\tilde{p}$ and impute the sample using the chosen method to obtain the imputed sample $x^*{}_{\text{imp}}$;
8: Obtain the value $T_n^*(x^*{}_{\text{imp}})$ of the BHEP statistic on the imputed dataset in the same way as in Step 3;
9: Repeat steps 4-5 B times to obtain the sequence of bootstrapped test statistics $T_{n,1}^*, \ldots, T_{n,B}^*$;
10: Reject the null hypothesis for the significance level α if $T_n(x)$ is greater than the $(1-\alpha)$-quantile of the empirical bootstrap distribution of $(T_{n,1}^*, \ldots, T_{n,B}^*)$.

presented in the Supplementary Material of Aleksić and Milošević (2025a). Similarly, here we present empirical tests' powers against standard Student's t distributions, while we omit the results for various scale matrices. Furthermore, the aforementioned Supplementary Material also contains the results for different columnwise missingness proportions.

Before we present the results of our study, we point to one of the most common misuses of the BHEP test in the presence of missing data, which consists of imputing the dataset but proceeding with the data analysis procedure that was originally designed only for complete samples. As illustrated in Figure 5.1, it is clear that, in the context of testing MVN, this is not a feasible approach, since the type I error is severely distorted.

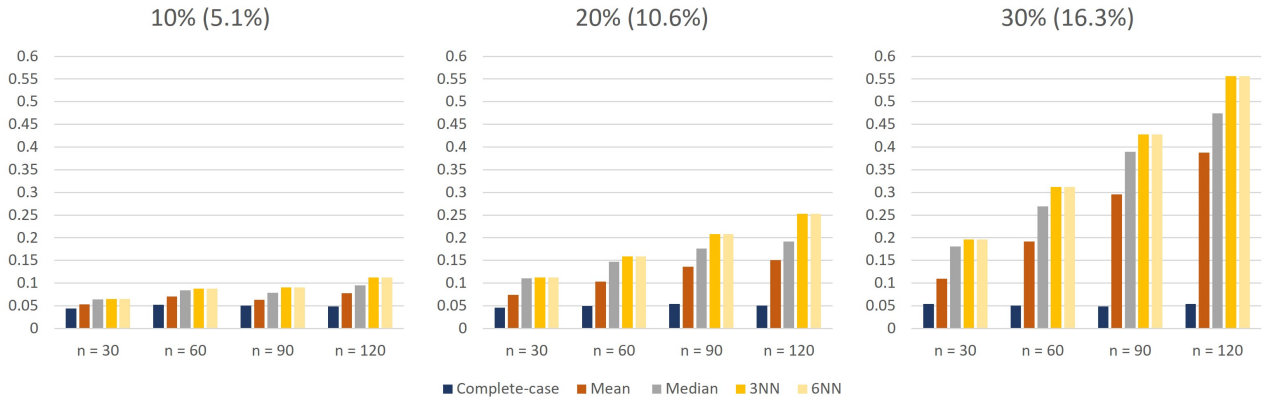

Figure 5.1: Empirical type I errors (y axis) for underlying 2-dimensional standard normal distribution and MCAR data, ignoring that the data were imputed (First percentage = probability that a row is incomplete, second percentage = probability that a value is missing).

Remark 5.3. One might initially find it counterintuitive that the distortion of type I error becomes more significant as the sample size increases. However, since the imputed values are not equally distributed as the available data, the test more accurately detects deviations from normality with larger sample sizes.

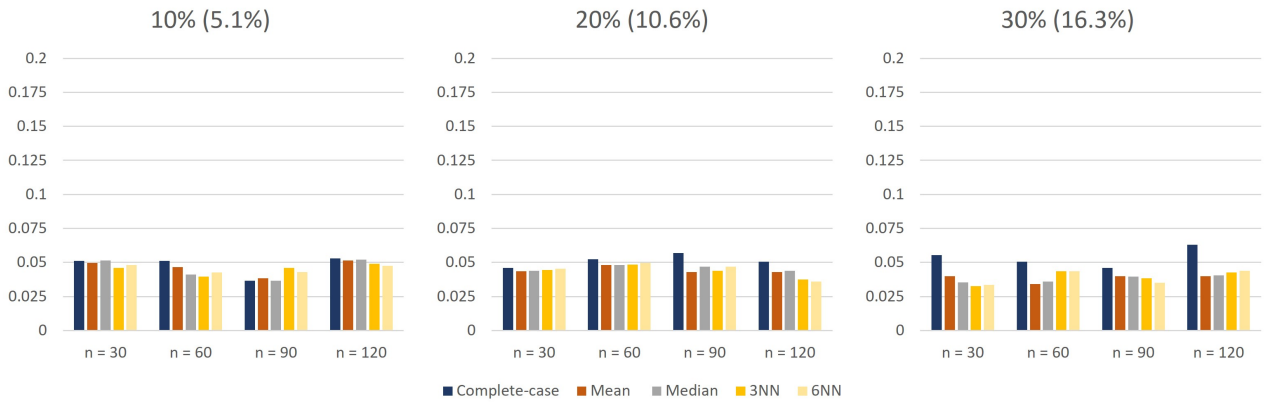

Figure 5.2: Empirical type I errors (y axis) for underlying 2-dimensional standard normal distribution and MCAR data (First percentage = probability that a row is incomplete, second percentage = probability that a value is missing).

As one can see from Figure 5.2, for the MCAR data and bivariate standard normal distribution, the complete-case approach presents itself as the best in terms of type I error preservation, and is, especially for moderate missingness, followed closely by the other methods. For the 3-dimensional case, however, as seen in Figure 5.3, kNN methods appear to be slightly more conservative and have empirical type I error further from the desired level 0.05. This becomes emphasized as the missingness probability starts to grow. However, in most of the real-world scenarios, where missingness is moderate, we can expect all of the methods to remain well-calibrated. Similar conclusions can be drawn for different correlation structures; for brevity, we omit those results.

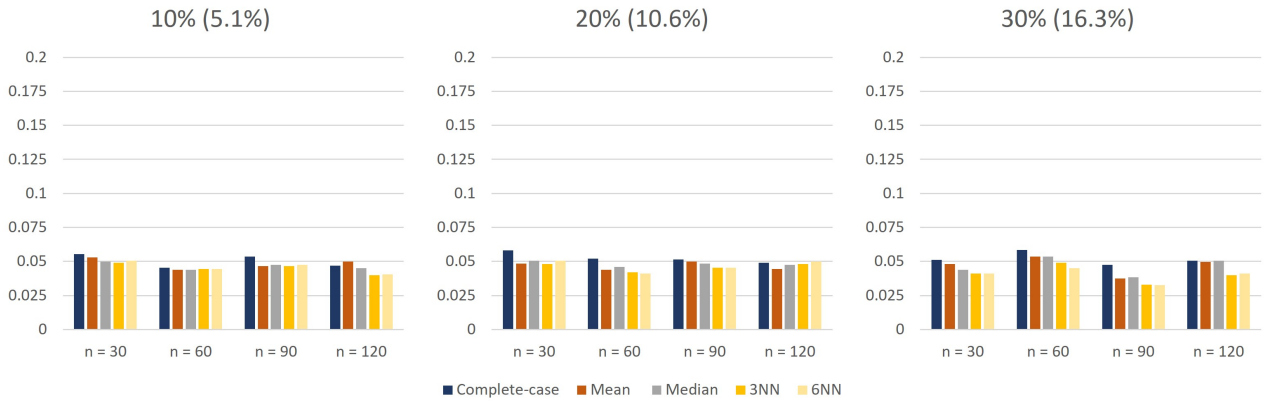

Figure 5.3: Empirical type I errors (y axis) for underlying 3-dimensional standard normal distribution and MCAR data (First percentage = probability that a row is incomplete, second percentage = probability that a value is missing).

This being said, we shift our focus to the power comparison. As can be seen from Figure 5.4, mean value imputation provides the greatest empirical powers, followed closely by median imputation. The kNN approaches significantly lag behind, while complete-case performs somewhere in the middle. One needs to point out that the advantage of the mean imputation approach is even more significant if we remember that empirical type I errors have shown that the complete-case approach has a higher tendency to reject the null hypothesis. The same relations are upheld for all of the other t-distributions, both standard and scaled, as well as for all of the observed dimensions. Those results also can be found in the Supplementary Material of Aleksić and Milošević (2025a).

Remark 5.4. Although it might initially seem unexpected that mean value and median impu-

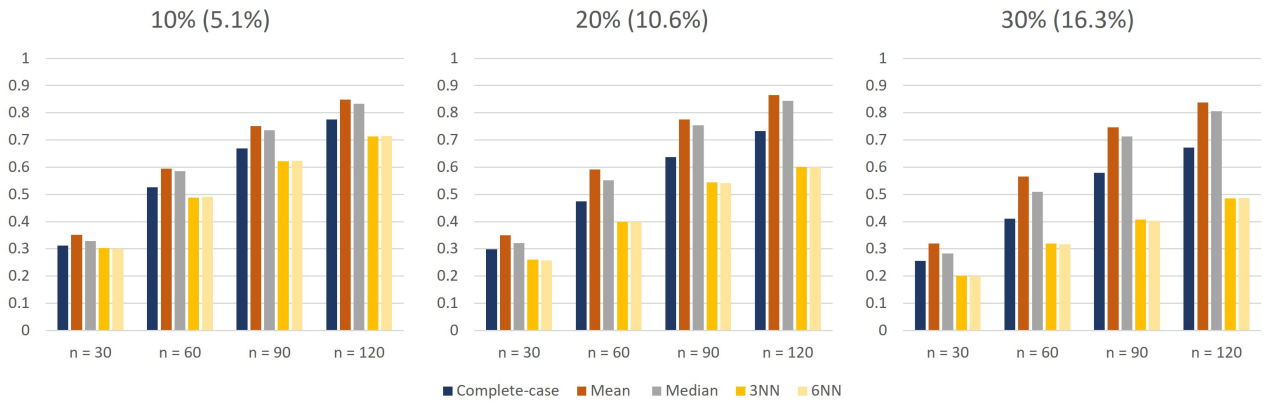

Figure 5.4: Empirical test powers (y axis) for underlying 2-dimensional t_5 distribution and MCAR data (First percentage = probability that a row is incomplete, second percentage = probability that a value is missing).

tation provide higher empirical power than the complete case, we believe this can be attributed to the ability of Algorithm 5.1 to utilize all of the partially observed data, unlike the complete-case approach. This effectively increases the sample size that the test works with.

In all of the studied scenarios, including those from the Supplementary Material of Aleksić and Milošević (2025a), Algorithm 5.1 combined with the mean imputation approach proved to be the best solution in terms of empirical power of the test, which is the primary advantage of this method. While median imputation offers a slightly less powerful approach, it still outperforms the complete-case. The main advantage of the complete-case method is its simplicity and its computational efficiency.

Once again, we emphasize that the MCAR assumption is necessary for the validity of both the complete-case method and Algorithm 5.1. Therefore, it is imperative to verify that the MCAR assumption holds before utilizing these approaches. This can be done using, for example, the well-known test developed by Little (1988) (having in mind that its validity relies on the normality of the data), or some of the recently developed tests, such as those by Aleksić (2024, 2025a), Berrett and Samworth (2023), or Bordino and Berrett (2024).

5.4 Real-data example

Before applying our methodology to a real-world dataset, we once again emphasize that, since our methodology is best suited for data that is missing according to the MCAR mechanism, it is imperative to verify that this assumption is met. For that purpose, we use two recently developed tests for MCAR, that do not require the normality assumption for their validity. In particular, we consider our covariance-based test from Section 3.3 (applied on centered data), and the compatibility-based test developed by Bordino and Berrett (2024). Both tests were used in Section 4.4. As it was mentioned there, the Little's MCAR test is sensitive to the MVN assumption, which we aim to test. Not to make a logical fallacy, we omit it from the MCAR testing here.

As an example demonstrating the ability of Algorithm 5.1 to detect departures from MVN, we once again consider the Sales and satisfaction dataset (Mahmoudi, 2024). A preliminary visual inspection of the variables, using histograms, density plots, or similar tools, suggests that it is of interest to examine whether the variables `Sales_Before` and `Sales_After` are jointly bivariate normal. Since the sample size in the original dataset is equal to 10000, due to the nature of real-world datasets in the context of GOF testing, it is expected that all of the testing procedures reject the null hypothesis with p-value equal to zero. This, indeed, is the

case for our dataset. Having this in mind, and knowing that maximum sample size covered in our simulation study was 120, a random subsample of that size was selected from the data, which resulted in missingness percentages of 18%, 15%, 17%, 5%, 13%, 22%, and 9% across all of the variables, in the same order as presented in Section 4.4.

Next, we test whether MCAR assumption holds for the selected sample. Although the MCAR assumption holds for the entire dataset, as seen in Section 4.4, for methodological purposes we conduct the testing once again for the selected subsample. Covariance-based test, applied to all variables, provides the p-value equal to 0.5. Compatibility-based test provides p-value equal to 1. When applied to the two variables of interest, the p-values are 0.72, and 0.98, respectively. We do not have enough evidence to reject MCAR, so we proceed with the MVN testing. The corresponding p-values can be seen in Table 5.1.

Table 5.1: Sales and satisfaction dataset: p-values for testing multivariate normality of `Sales_Before` and `Sales_After` on the subsample of size 120.

Complete-case	Mean	Median	3NN	6NN
0	0.002	0.002	0.112	0.112

As observed, the complete-case approach, as well as mean and median imputation, are able to detect departures from bivariate normality at the standard 5% significance level for this sample. In contrast, imputation using 3 and 5 nearest neighbors produces substantially higher p-values, resulting in a failure to reject MVN. This difference can be attributed to the markedly lower power of the $3NN$ and $5NN$ methods, as demonstrated in the power study.

Chapter 6

Multivariate two-sample hypothesis testing in the presence of missing data

Testing whether two samples originate from the same probability distribution, known as two-sample hypothesis testing, is a fundamental problem in statistical theory with a broad spectrum of applications across various fields. In medical research, such tests are used to compare patient outcomes between treatment and control groups, assessing the effectiveness of new drugs or interventions. For example, in quality control, manufacturers employ these methods to verify whether changes in production processes affect product characteristics. Environmental scientists utilize them to assess changes in climate variables, such as temperature distributions over time, while geneticists apply them to compare gene expression profiles between different populations.

Given the versatility of possible applications, two-sample tests continue to play a crucial role in data-driven decision-making across numerous disciplines, and, consequently, they have been extensively studied in the literature. Here, we present some notable examples. The idea of testing whether two samples originate from the same probability distribution first appeared for one-dimensional data, and can be traced back to Pearson (1900), who introduced the well-known χ^2-test. Later, in the early 20th century, Student (1908) proposed a test for differences between normal distributions by comparing the means; the idea was later generalized by Fisher (1925) in his works on Analysis of Variance (ANOVA). These methods are parametric, and they have severe limitations in non-normal settings.

Early nonparametric solutions for this problem were presented in the form of the well-known Kolmogorov-Smirnov (KS) test, which was introduced by Kolmogorov (1933) and generalized by Smirnov (1939). The test is based on measuring the discrepancy of ECDFs between the two samples. The Wilcoxon–Mann–Whitney test (Wilcoxon, 1945; Mann and Whitney, 1947) is another nonparametric two-sample test for one-dimensional data that was based on comparing the rank-sums between groups.

Cramér (1928) and von Mises (1928) considered the GOF testing by comparing the integrated square distance between the CDFs. Anderson (1962) later extended the idea to the two-sample setting. Another approach that utilized integrated square difference of ECDFs came with Pettitt (1976), who modified the test statistic of the GOF test by Anderson and Darling (1954) for the two-sample problem.

The test proposed by by Hotelling (1931) can be considered to be one of the first two-sample tests for multivariate data. The test was based on a test statistic that generalizes the Student's ratio, and, as such, it relies on the normality assumption, as well as on the assumption of equal covariances. Notable example of a nonparametric multivariate two-sample test is the test by Friedman and Rafsky (1979) that generalizes the two-sample runs test by Wald and Wolfowitz (1940). Some of the modern examples are the energy distance-based two-sample test (Bar-

inghaus and Franz, 2004; Székely and Rizzo, 2004), the Maximum Mean Discrepancy (MMD) kernel-based approach by Gretton et al. (2012) (their equivalence is shown by Sejdinovic et al. (2013)), as well as the binary classifier-based approach by Lopez-Paz and Oquab (2016). For count data data, a useful reference is Alba-Fernández et al. (2017), and for matrix data we recommend Lukić and Milošević (2024).

Results obtained from real-world data indicate that complex data types, like audio and images, despite their high dimensionality, tend to be structured around a lower-dimensional manifold. Due to this fact, the standard energy distance-based two-sample test, which is built upon Euclidean distance, performs well for low-dimensional data, and loses power quickly for data in high dimensional space (e.g. Chu and Dai, 2024, and references therein). Consequently, much work has been done in that direction: to modify the original test to be able to recognize the inner structure of the data and effectively reduce the dimension. A nice overview can be found in a paper by Chu and Dai (2024). However, in terms of adapting the original test to work with incomplete samples, not much progress has been made. Knowing the importance and the wide applicability of the test, it is essential to address that issue. Our findings, that are presented in this chapter, aim to begin filling this gap in the literature.

In this chapter, we consider the problem of two-sample testing in the presence of missing data under the broad class of missingness mechanisms, presenting the results of Aleksić and Milošević (2025b). For this purpose, we focus on the energy-based two-sample test (Baringhaus and Franz, 2004; Székely and Rizzo, 2004). Besides the complete-case approach for handling missing data, we propose a novel modification of the test statistic that utilizes all available data, along with two resampling procedures for approximating the corresponding p-values. A novel bootstrap method is also introduced for p-value approximation when the test statistic is computed on samples filled using commonly used imputation methods for handling missing data. In an extensive simulation study, all approaches are compared in terms of preservation of type I error and in terms of power. General recommendations are given for each of the studied scenarios.

The results are presented according to the following structure. In Section 6.1, we restate some basic properties of the original energy test that are necessary for our further research. In Section 6.2, we introduce our novel testing procedures and outline their expected strengths and flaws. Some theoretical results regarding the null distribution of the novel test are also provided. Section 6.3 is devoted to the extensive simulation study, with the aim of examining the performance of novel procedures in terms of preservation of type I error and empirical power. Depending on the specific scenario, such as data distribution and the underlying missingness mechanism, certain recommendations are provided at the end of the chapter.

6.1 Revisiting the energy test

Let $\mathscr{M}$ be a metric space and let μ and ν be two probability measures on it. Assume we have two independent samples of IID random elements in $\mathscr{M}$:

$$X_1, \ldots, X_n \sim \mu \quad \text{and} \quad Y_1, \ldots, Y_m \sim \nu. \tag{6.1}$$

A commonly tested hypothesis in a two-sample test is whether the two samples originate from the same distribution

$$H_0 : \mu = \nu, \tag{6.2}$$

against the complementary alternative. If ρ is a metric on $\mathscr{M}$, one can define the *energy distance* between the distributions μ and ν as

$$D(\mu, \nu) = 2\mathbb{E}\rho(X, Y) - \mathbb{E}\rho(X, X') - \mathbb{E}\rho(Y, Y'), \tag{6.3}$$

where $X, X' \sim \mu$ and $Y, Y' \sim \nu$ are all mutually independent.

The name *energy distance* comes from physics. In physics, particularly in electrostatics, gravity, or molecular systems, we often deal with systems of many particles that interact pairwise (e.g., via gravitational or electric forces). In such a system, we can distinguish: (i) energy due to interactions within a group (e.g., within one cloud of particles), and (ii) energy due to interactions between particles from different groups. Physicists are often interested in the *net interaction energy*, which is the difference between the total interaction energy and the internal self-energies.

The term $\mathbb{E}\rho(X, Y)$ in (6.3) can be interpreted as the total interaction energy between μ and ν, i.e. the average distance between the two randomly chosen points from them. Similarly, $\mathbb{E}\rho(X, X')$ ($\mathbb{E}\rho(Y, Y')$) can be interpreted as the internal energy of μ (ν), the average distance between the two randomly chosen points from μ (ν). So, when computing the energy distance $D(\mu, \nu)$, we are basically measuring the *excess interaction energy* between μ and ν, beyond the energy one would expect if both samples were from the same distribution.

To further clarify this interpretation, one may ask: Is the separation between the distributions greater than what would be expected solely due to their natural variability? If the answer is negative, the energy distance is expected to be close to zero, and to deviate significantly from zero otherwise. For additional parallels to physics and further motivation, we recommend consulting the paper by Székely and Rizzo (2013).

The test statistic of the energy test (Baringhaus and Franz, 2004; Székely and Rizzo, 2004) is based on the sample estimate of $D(\mu, \nu)$ and is defined as

$$T_{nm} = \frac{2}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}\rho(X_i, Y_j) - \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\rho(X_i, X_j) - \frac{1}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m}\rho(Y_i, Y_j). \tag{6.4}$$

The energy distance is not always the proper metric on the set of all probability distributions on $\mathscr{M}$. Sometimes it is not positive; a sufficient condition, for example, is for $(\mathscr{M}, \rho)$ to be a metric space of a *strong negative type*. Specifically, $(\mathscr{M}, \rho)$ is of a strong negative type if for any two probability measures μ and ν, with finite first moments, it holds that

$$\iint \rho(u, v)\,\mathrm{d}\mu(u)\,\mathrm{d}\mu(v) + \iint \rho(u, v)\,\mathrm{d}\nu(u)\,\mathrm{d}\nu(v) - 2\iint \rho(u, v)\,\mathrm{d}\mu(u)\,\mathrm{d}\nu(v) \leqslant 0,$$

and the left hand side is equal to zero if and only if $\mu = \nu$. As a consequence, on a space of strong negative type, the energy distance is able to differentiate between any two probability distributions. An example of one such space, that is of key interest in mathematical statistics is any separable Hilbert space. As a special case, we have the Euclidean space $(\mathbb{R}^d, \|\cdot\|)$ with standard Euclidean metric. Since that space is of our main interest, further discussion would fall out from the scope of this paper. For more details, one should consult, e.g., Klebanov et al. (2005), Chu and Dai (2024), and references therein.

6.2 Novel procedures

In this section, we propose several novel adaptations of the energy test suitable for incomplete samples with missingness mechanisms that are not necessarily MCAR. The main assumption is that the samples in (6.1) are independent, and that both consist of IID random vectors from $\mathbb{R}^d$. Generally speaking, our methodology can be applied to any metric ρ on $\mathbb{R}^d$, although our main focus will be on the standard Euclidean metric, which will be thoroughly examined in the extensive simulation study that will follow, where the novel approaches will be compared to the complete-case analysis, which is used as a benchmark.

First of all, let us introduce some basic notation that we will need, which is similar to the one used in previous chapters. For every element X_i of the sample $X_1,\dots,X_n$ let R_i^X be the corresponding vector of the same length d as X_i, whose kth element is equal to 1 if the kth element of X_i is observed, and 0 otherwise. We will refer to R_i^X as a *response indicator vector* of X_i. Let S_i^X be the indicator that each component of X_i is observed, i.e. indicator that X_i is the *complete case*. Let R_i^Y and S_i^Y be defined in a similar manner.

The number of complete cases from the sample $X_1,\dots,X_n$ will be denoted as $\hat{n}$; it is clear that $\hat{n}=\sum_{i=1}^n S_i^X$. Similarly, let $\hat{m}=\sum_{j=1}^m S_j^Y$ be the number of complete cases from the sample $Y_1,\dots,Y_m$. Finally, let $\odot$ denote the standard Hadamard–Schur componentwise multiplication:

$$(u_1,u_2,\dots,u_d)\odot(v_1,v_2,\dots,v_d)=(u_1v_1,u_2v_2,\dots,u_dv_d).$$

6.2.1 Complete-case analysis: the benchmark

First, and fairly common approach for handling missing data in practice is the complete-case analysis, where the test statistic T_{nm} is calculated only on the completely observed sample elements. In our notation, the statistic can be written as

$$T_{nm}^{CC}=\frac{2}{\hat{n}\hat{m}}\sum_{i=1}^n\sum_{j=1}^m\rho(X_i,Y_j)S_i^XS_j^Y-\frac{1}{\hat{n}^2}\sum_{i=1}^n\sum_{j=1}^n\rho(X_i,X_j)S_i^XS_j^X-\frac{1}{\hat{m}^2}\sum_{i=1}^n\sum_{j=1}^n\rho(Y_i,Y_j)S_i^YS_j^Y. \quad (6.5)$$

The following theorem states that, as expected, the complete-case test statistic has the same asymptotic distribution as the complete-sample one, under the assumption of MCAR data. This result might seem obvious, but results such as those from Chapters 4 and 5 demonstrate that the formal proofs can be very challenging.

Theorem 6.1 [Aleksić, Milošević (2025)]**.** *Let $X_1,X_2,\dots,X_n$ and $Y_1,Y_2,\dots,Y_m$ be two samples of random vectors from $\mathbb{R}^d$, such that the variables $X_1,\dots,X_n,Y_1,\dots,Y_m$ are independent and identically distributed with characteristic function $\varphi(t)$. Let T_{nm} and T_{nm}^{CC} be as in* (6.4) *and* (6.5)*, respectively. Finally, assume that both samples have equal probabilities of a case being complete, i.e. $\mathrm{E}S^X=\mathrm{E}S^Y=q$. Then, under MCAR, it holds that $\frac{nm}{n+m}T_{nm}$ and $\frac{\hat{n}\hat{m}}{\hat{n}+\hat{m}}T_{nm}^{CC}$ have the same asymptotic distribution. More precisely,*

$$\frac{nm}{n+m}T_{nm}\xrightarrow{D}\|Z(t)\|_w^2,\quad \frac{\hat{n}\hat{m}}{\hat{n}+\hat{m}}T_{nm}^{CC}\xrightarrow{D}\|Z(t)\|_w^2,\qquad as\quad n,m\to\infty,\quad \frac{n}{n+m}\to\lambda^2,$$

where $\{Z(t)\mid t\in\mathbb{R}^d\}$ is a centered Gaussian process with covariance operator defined for $s,t\in\mathbb{R}^d$ as

$$\begin{aligned}C(t,s)&=\mathbb{E}(Z(s)Z(t))\\&=\mathrm{Re}\big(\varphi(s-t)\big)+\mathrm{Im}\big(\varphi(s+t)\big)\\&\quad-\mathrm{Re}\big(\varphi(t)\big)\mathrm{Re}\big(\varphi(s)\big)-\mathrm{Im}\big(\varphi(t)\big)\mathrm{Re}\big(\varphi(s)\big)-\mathrm{Im}\big(\varphi(s)\big)\mathrm{Re}\big(\varphi(t)\big)-\mathrm{Im}\big(\varphi(s)\big)\mathrm{Im}\big(\varphi(t)\big).\end{aligned} \quad (6.6)$$

Here, $\langle f,g\rangle_w=\int_{\mathbb{R}^d}f(t)g(t)w(t)\mathrm{d}t$, and $\|f\|_w^2=\langle f,f\rangle_w$, where $w(t)=\|t\|^{d-1}$.

Proof. We adapt the proof of Theorem 1 from Chen et al. (2019), where they treated the problem of convergence of T_{nn}, i.e. the case where $m=n$.

It is a known result (see, e.g., Chen et al., 2019) that

$$T_{mn}=\int_{\mathbb{R}^d}|\varphi_n(t)-\varphi_m(t)|^2w(t)\mathrm{d}t, \quad (6.7)$$

where $w(t) = \|t\|^{d-1}$ and $\varphi_n(t) = \frac{1}{n}\sum_{i=1}^{n} e^{it^T X_i}$ is the empirical characteristic function of the sample $X_1, X_2, \ldots, X_m$ and φ_m is defined similarly.

It is now readily seen that

$$
\begin{aligned}
\frac{\hat{n}\hat{m}}{\hat{n}+\hat{m}} T_{nm}^{CC} &= \frac{\hat{n}\hat{m}}{\hat{n}+\hat{m}} \int_{\mathbb{R}^d} \left| \frac{1}{\hat{n}} \sum_{k=1}^{n} e^{it^T X_k} S_k^X - \frac{1}{\hat{m}} \sum_{j=1}^{m} e^{it^T Y_j} S_j^Y \right|^2 w(t)\mathrm{d}t \\
&= \frac{\hat{n}\hat{m}}{\hat{n}+\hat{m}} \int_{\mathbb{R}^d} \left| \left[\frac{1}{\hat{n}} \sum_{k=1}^{n} \cos(t^T X_k) S_k^X - \frac{1}{\hat{m}} \sum_{j=1}^{m} \cos(t^T Y_j) S_j^Y \right] \right. \\
&\qquad \left. + i \left[\frac{1}{\hat{n}} \sum_{k=1}^{n} \sin(t^T X_k) S_k^X - \frac{1}{\hat{m}} \sum_{j=1}^{m} \sin(t^T Y_j) S_j^Y \right] \right|^2 w(t)\mathrm{d}t \\
&= \frac{\hat{n}\hat{m}}{\hat{n}+\hat{m}} \int_{\mathbb{R}^d} \left(\left[\frac{1}{\hat{n}} \sum_{k=1}^{n} \cos(t^T X_k) S_k^X - \frac{1}{\hat{m}} \sum_{j=1}^{m} \cos(t^T Y_j) S_j^Y \right]^2 \right. \\
&\qquad \left. + \left[\frac{1}{\hat{n}} \sum_{k=1}^{n} \sin(t^T X_k) S_k^X - \frac{1}{\hat{m}} \sum_{j=1}^{m} \sin(t^T Y_j) S_j^Y \right]^2 \right) w(t)\mathrm{d}t \\
&= \frac{\hat{n}\hat{m}}{\hat{n}+\hat{m}} \int_{\mathbb{R}^d} \left[\frac{1}{\hat{n}} \sum_{k=1}^{n} \cos(t^T X_k) S_k^X - \frac{1}{\hat{m}} \sum_{j=1}^{m} \cos(t^T Y_j) S_j^Y \right. \\
&\qquad \left. + \frac{1}{\hat{n}} \sum_{k=1}^{n} \sin(t^T X_k) S_k^X - \frac{1}{\hat{m}} \sum_{j=1}^{m} \sin(t^T Y_j) S_j^Y \right]^2 w(t)\mathrm{d}t \\
&= \frac{\hat{n}\hat{m}}{\hat{n}+\hat{m}} \int_{\mathbb{R}^d} \left[\frac{1}{\hat{n}} \sum_{k=1}^{n} \big(\cos(t^T X_k) + \sin(t^T X_k)\big) S_k^X - \frac{1}{\hat{m}} \sum_{j=1}^{m} \big(\cos(t^T Y_j) + \sin(t^T Y_j)\big) S_j^Y \right]^2 w(t)\mathrm{d}t \\
&= \frac{\hat{n}\hat{m}}{\hat{n}+\hat{m}} \int_{\mathbb{R}^d} \left[\frac{1}{\hat{n}} \sum_{k=1}^{n} \Big[\big(\cos(t^T X_k) + \sin(t^T X_k) - \big(\mathrm{Re}(\varphi(t)) + \mathrm{Im}(\varphi(t))\big)\big) S_k^X \Big] \right. \\
&\qquad + \frac{1}{\hat{n}} \sum_{k=1}^{n} \big(\mathrm{Re}(\varphi(t)) + \mathrm{Im}(\varphi(t))\big) S_k^X \\
&\qquad - \frac{1}{\hat{m}} \sum_{j=1}^{m} \Big[\big(\cos(t^T Y_j) + \sin(t^T Y_j) - \big(\mathrm{Re}(\varphi(t)) + \mathrm{Im}(\varphi(t))\big)\big) S_j^Y \Big] \\
&\qquad \left. - \frac{1}{\hat{m}} \sum_{j=1}^{m} \big(\mathrm{Re}(\varphi(t)) + \mathrm{Im}(\varphi(t))\big) S_j^Y \right]^2 w(t)\mathrm{d}t \\
&= \frac{\hat{n}\hat{m}}{\hat{n}+\hat{m}} \int_{\mathbb{R}^d} \left[\frac{1}{\hat{n}} \sum_{k=1}^{n} \Big[\big(\cos(t^T X_k) + \sin(t^T X_k) - \big(\mathrm{Re}(\varphi(t)) + \mathrm{Im}(\varphi(t))\big)\big) S_k^X \Big] \right. \\
&\qquad - \frac{1}{\hat{m}} \sum_{j=1}^{m} \Big[\big(\cos(t^T Y_j) + \sin(t^T Y_j) - \big(\mathrm{Re}(\varphi(t)) + \mathrm{Im}(\varphi(t))\big)\big) S_j^Y \Big] \\
&\qquad \left. + \frac{1}{\hat{n}} \sum_{k=1}^{n} \big(\mathrm{Re}(\varphi(t)) + \mathrm{Im}(\varphi(t))\big) S_k^X - \frac{1}{\hat{m}} \sum_{j=1}^{m} \big(\mathrm{Re}(\varphi(t)) + \mathrm{Im}(\varphi(t))\big) S_j^Y \right]^2 w(t)\mathrm{d}t
\end{aligned}
$$

$$
\begin{aligned}
&= \int_{\mathbb{R}^d} \Bigg[\sqrt{\frac{\hat{m}}{\hat{n}+\hat{m}}} \frac{1}{\sqrt{\hat{n}}} \sum_{k=1}^{n} \Big[\big(\cos(t^T X_k) + \sin(t^T X_k) - \big(\mathrm{Re}(\varphi(t)) + \mathrm{Im}(\varphi(t))\big)\big) S_k^X \Big] \\
&\qquad - \sqrt{\frac{\hat{n}}{\hat{n}+\hat{m}}} \frac{1}{\sqrt{\hat{m}}} \sum_{j=1}^{m} \Big[\big(\cos(t^T Y_j) + \sin(t^T Y_j) - \big(\mathrm{Re}(\varphi(t)) + \mathrm{Im}(\varphi(t))\big)\big) S_j^Y \Big] \\
&\quad + \underbrace{\sqrt{\frac{\hat{n}\hat{m}}{\hat{n}+\hat{m}}} \big(\mathrm{Re}(\varphi(t)) + \mathrm{Im}(\varphi(t))\big) \frac{1}{\hat{n}} \underbrace{\sum_{k=1}^{n} S_k^X}_{=\hat{n}} - \sqrt{\frac{\hat{n}\hat{m}}{\hat{n}+\hat{m}}} \big(\mathrm{Re}(\varphi(t)) + \mathrm{Im}(\varphi(t))\big) \frac{1}{\hat{m}} \underbrace{\sum_{j=1}^{m} S_j^Y}_{=\hat{m}}}_{=0} \Bigg]^2 w(t)\,\mathrm{d}t \\
&= \int_{\mathbb{R}^d} \Bigg[\sqrt{\frac{\hat{m}}{\hat{n}+\hat{m}}} \frac{1}{\sqrt{\hat{n}}} \sum_{k=1}^{n} \Big[\big(\cos(t^T X_k) + \sin(t^T X_k)\big) - \big(\mathrm{Re}(\varphi(t)) + \mathrm{Im}(\varphi(t))\big) \Big] S_k^X \\
&\qquad - \sqrt{\frac{\hat{n}}{\hat{n}+\hat{m}}} \frac{1}{\sqrt{\hat{m}}} \sum_{j=1}^{m} \Big[\big(\cos(t^T Y_j) + \sin(t^T Y_j)\big) - \big(\mathrm{Re}(\varphi(t)) + \mathrm{Im}(\varphi(t))\big) \Big] S_j^Y \Bigg]^2 w(t)\,\mathrm{d}t \\
&= \int_{\mathbb{R}^d} \Bigg[\frac{\sqrt{\frac{\hat{m}}{\hat{n}+\hat{m}}} \frac{1}{\sqrt{\hat{n}}}}{\sqrt{\frac{m}{n+m}} \frac{1}{\sqrt{n}}} \sqrt{q} \sqrt{\frac{m}{n+m}} \frac{1}{\sqrt{n}} \sum_{k=1}^{n} \Big[\big(\cos(t^T X_k) + \sin(t^T X_k)\big) - \big(\mathrm{Re}(\varphi(t)) + \mathrm{Im}(\varphi(t))\big) \Big] \frac{S_k^X}{\sqrt{q}} \\
&\qquad - \frac{\sqrt{\frac{\hat{n}}{\hat{n}+\hat{m}}} \frac{1}{\sqrt{\hat{m}}}}{\sqrt{\frac{n}{n+m}} \frac{1}{\sqrt{m}}} \sqrt{q} \sqrt{\frac{n}{n+m}} \frac{1}{\sqrt{m}} \sum_{j=1}^{m} \Big[\big(\cos(t^T Y_j) + \sin(t^T Y_j)\big) \\
&\qquad\qquad - \big(\mathrm{Re}(\varphi(t)) + \mathrm{Im}(\varphi(t))\big) \Big] \frac{S_j^Y}{\sqrt{q}} \Bigg]^2 w(t)\,\mathrm{d}t. \qquad (6.8)
\end{aligned}
$$

Now, by the Law of Large Numbers and Continuous Mapping Theorem we have that

$$
\frac{\sqrt{\frac{\hat{m}}{\hat{n}+\hat{m}}} \frac{1}{\sqrt{\hat{n}}}}{\sqrt{\frac{m}{n+m}} \frac{1}{\sqrt{n}}} \sqrt{q} \xrightarrow{P} 1, \quad \text{and} \quad \frac{\sqrt{\frac{\hat{n}}{\hat{n}+\hat{m}}} \frac{1}{\sqrt{\hat{m}}}}{\sqrt{\frac{n}{n+m}} \frac{1}{\sqrt{m}}} \sqrt{q} \xrightarrow{P} 1. \qquad (6.9)
$$

By Slutsky's theorem, these terms can be asymptotically treated as 1.

Denote

$$
Z_{n,1}(t) := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \Big[\big(\cos(t^T X_i) + \sin(t^T X_i)\big) - \big(\mathrm{Re}(\varphi(t)) + \mathrm{Im}(\varphi(t))\big) \Big] \frac{S_i^X}{\sqrt{q}}
$$

and

$$
Z_{m,2}(t) := \frac{1}{\sqrt{m}} \sum_{j=1}^{m} \Big[\big(\cos(t^T Y_j) + \sin(t^T Y_j)\big) - \big(\mathrm{Re}(\varphi(t)) + \mathrm{Im}(\varphi(t))\big) \Big] \frac{S_j^Y}{\sqrt{q}}.
$$

One can easily note that for every $1 \leqslant i \leqslant n$ and $1 \leqslant j \leqslant n$ the random functions

$$
h((X_i, R_i^X), t) := \Big[\big(\cos(t^T X_i) + \sin(t^T X_i)\big) - \big(\mathrm{Re}(\varphi(t)) + \mathrm{Im}(\varphi(t))\big) \Big] \frac{S_i^X}{\sqrt{q}}
$$

and

$$
h((Y_j, R_j^Y), t) := \Big[\big(\cos(t^T Y_j) + \sin(t^T Y_j)\big) - \big(\mathrm{Re}(\varphi(t)) + \mathrm{Im}(\varphi(t))\big) \Big] \frac{S_j^Y}{\sqrt{q}}
$$

are independent and identically distributed centered random elements of $L^2(\mathbb{R}^d, w(t)\mathrm{d}t)$ with covariance function equal to

$$
\begin{aligned}
&\mathrm{E}\Big\{\Big[\big(\cos(t^T X_1)+\sin(t^T X_1)\big)-\big(\mathrm{Re}(\varphi(t))+\mathrm{Im}(\varphi(t))\big)\Big] \\
&\qquad\cdot\Big[\big(\cos(s^T X_1)+\sin(s^T X_1)\big)-\big(\mathrm{Re}(\varphi(s))+\mathrm{Im}(\varphi(s))\big)\Big]\frac{(S_1^X)^2}{q}\Big\} \\
&=\mathrm{E}\big[\cos(t^T X_1)\cos(s^T X_1)+\cos(t^T X_1)\sin(s^T X_1)+\sin(t^T X_1)\cos(s^T X_1)+\sin(t^T X_1)\sin(s^T X_1)\big] \\
&\qquad-\big(\mathrm{Re}(\varphi(t))+\mathrm{Im}(\varphi(t))\big)\cdot\big(\mathrm{Re}(\varphi(s))+\mathrm{Im}(\varphi(s))\big) \\
&=\mathrm{E}\Big[\frac{1}{2}\Big(\cos\big((t-s)^T X_1\big)+\cos\big(((t+s)^T X_1\big)\Big)+\frac{1}{2}\Big(\sin\big((s-t)^T X_1\big)+\sin\big((s+t)^T X_1\big)\Big) \\
&\qquad+\frac{1}{2}\Big(\sin\big((t-s)^T X_1\big)+\sin\big((s+t)^T X_1\big)\Big)+\frac{1}{2}\Big(\cos\big((t-s)^T X_1\big)-\cos\big((t+s)^T X_1\big)\Big)\Big] \\
&\qquad-\big(\mathrm{Re}(\varphi(t))+\mathrm{Im}(\varphi(t))\big)\cdot\big(\mathrm{Re}(\varphi(s))+\mathrm{Im}(\varphi(s))\big) \\
&=\mathrm{Re}(\varphi(t-s))+\mathrm{Im}(\varphi(t+s))-\mathrm{Re}(\varphi(t))\mathrm{Re}(\varphi(s))-\mathrm{Im}(\varphi(t))\mathrm{Re}(\varphi(s)) \\
&\qquad-\mathrm{Re}(\varphi(t))\mathrm{Im}(\varphi(s))-\mathrm{Im}(\varphi(t))\mathrm{Im}(\varphi(s)),
\end{aligned}
\tag{6.10}
$$

which is exactly the covariance function $C(t,s)$ from (6.6).

Now, since the random functions $h((X_i, R_i^X), t)$ and $h((Y_j, R_j^Y), t)$ are elements of $L^2(\mathbb{R}^d, w(t)\mathrm{d}t)$ with existing covariance function (6.10), we can apply the Central Limit Theorem for Hilbert spaces (e.g., Henze, 2024, Theorem 17.29) to conclude that

$$
\sqrt{\frac{m}{n+m}}Z_{n,1}(t)\xrightarrow{D}Z_1(t)\quad\text{and}\quad\sqrt{\frac{n}{n+m}}Z_{m,2}(t)\xrightarrow{D}Z_2(t),
$$

where $\{Z_1(t)\mid t\in\mathbb{R}^d\}$ and $\{Z_2(t)\mid t\in\mathbb{R}^d\}$ are independent, centered Gaussian processes with covariance functions equal to $\lambda^2 C(t,s)$ and $(1-\lambda^2)C(t,s)$, respectively, where $C(t,s)$ is defined in (6.6). Having the independence of Z_1 and Z_2 and the convergence (6.9) we can conclude that the difference inside the large square brackets in (6.8) converges in distribution to the random process

$$
Z(t)=Z_1(t)-Z_2(t),
$$

with covariance function equal, due to independence, to the sum of the corresponding covariance functions:

$$
\lambda^2 C(t,s)+(1-\lambda^2)C(t,s)=C(t,s).
$$

Recalling the definition (and continuity) of $\|\cdot\|_w$, we finally conclude that

$$
\frac{\hat{n}\hat{m}}{\hat{n}+\hat{m}}T_{nm}^{CC}\xrightarrow{D}\|Z(t)\|_w^2,
$$

where $\{Z(t)\mid t\in\mathbb{R}^d\}$ is the centered Gaussian process with the covariance function $C(t,s)$ given in (6.6), which concludes this part of the proof.

The proof that $nmT_{nm}/(n+m)$ also converges to $\|Z(t)\|_w^2$ follows directly from the fact that $\hat{n}=n$, $\hat{m}=m$ and $q=1$, and is known from the literature (see Chen et al., 2019). This concludes the proof of Theorem 6.1. ■

As in the original (complete-sample) energy test, the asymptotic null distribution of the test statistic depends on the underlying distribution of the data and is therefore not distribution-free, so the bootstrap algorithm is utilized for calculating the critical values or p-values of the test. The two proposed resampling procedures will be presented in the Subsection 6.2.4.

Due to its simplicity and low computational cost, complete-case analysis became one of the most commonly used approaches of handling missing data when conducting various statistical analyses on incomplete datasets. Generally speaking, it can serve as a quick solution when data are MCAR, or missingness rate is very low. Furthermore, for some procedures, such as independence testing (Aleksić et al., 2023), or MVN testing (Aleksić and Milošević, 2025a), complete-case analysis performed better under MCAR than some imputation methods. However, when either data are not MCAR, or the missingness rate is very high, complete-case is known to leave much to be desired, producing biased estimates, decreasing the power of the test, having no type I error control, and exibiting other limitations (e.g., Aleksić and Milošević, 2025a; Tsatsi et al., 2024, and others). This is especially noticeable when dealing with high-dimensional data, where restricting analysis to complete cases can, obviously, result in a significant loss of information and statistical efficiency. The energy test is not an exception to this rule. For example, simulations by Zeng et al. (2024) indicate that, under their specific *MMD-Miss* approach and MNAR setting, complete-case analysis exibits type I error asymptotically equal to 1.

Given the aforementioned flaws of complete-case analysis, i.e. wastefulness and sensitivity to data not being MCAR, it is imperative to seek for better approaches for conducting the energy test on incomplete samples, which could, at least partially, overcome these limitations.

As our simulations will demonstrate, complete case analysis can, in many scenarios, including the MCAR setting, be outperformed by certain weighting and imputation methods, when appropriate bootstrap resampling is employed.

6.2.2 Weighting methods

As we have discussed, partially observed cases, although incomplete, may still carry useful information about the underlying structure of the data, or parameters of interest, and should not be disregarded outright. A natural way to incorporate these cases is to assign them weights based on the amount of observed information they contain. This approach ensures that observations with more observed components have a proportionally greater impact on the test statistic, hence making fuller use of the available data while acknowledging varying degrees of completeness across cases. For that purpose, we modify the original (de facto Euclidean) distance into a *weighted distance* as

$$\rho_W\big((X,R^X),(Y,R^Y)\big)=\rho_{\text{trunc}}\big((X,R^X),(Y,R^Y)\big)\cdot\frac{\sum_{i=1}^{d}(R^X\odot R^Y)_i}{d}, \tag{6.11}$$

where $(R^X\odot R^Y)_i$ is the ith component of $R^X\odot R^Y$, and distance $\rho_{\text{trunc}}\big((X,R^X),(Y,R^Y)\big)$ is calculated between those subvectors of X and Y that are observed in both cases. The weight $\frac{1}{d}\sum_{i=1}^{d}(R^X\odot R^Y)_i$ is assigned such that that complete cases receive a weight of 1, while cases with more missing components receive smaller weights, contributing proportionally less to the overall sum. Naturally, the *weighted test statistic* is defined as

$$\begin{aligned}T_{nm}^{W}=\frac{2}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}\rho_W\big((X_i,R_i^X),(Y_j,R_j^Y)\big)-\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\rho_W\big((X_i,R_i^X),(X_j,R_j^X)\big)\\-\frac{1}{m^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\rho_W\big((Y_i,R_i^Y),(Y_j,R_j^Y)\big).\end{aligned} \tag{6.12}$$

Unlike for the statistics T_{nm} and T_{nm}^{CC}, there is no representation analogous to (6.7) for statistic T_{nm}^{W}. An alternative approach would be to express it as

$$T_{nm}^W = \frac{1}{n^2m^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{m}\sum_{l=1}^{m}\Big(\rho_W\big((X_i,R_i^X),(Y_l,R_l^Y)\big)+\rho_W\big((X_j,R_j^X),(Y_k,R_k^Y)\big) \\ -\rho_W\big((X_i,R_i^X),(X_j,R_j^X)\big)-\rho_W\big((Y_k,R_k^Y),(Y_l,R_l^Y)\big)\Big) \\ =: \frac{1}{n^2m^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{m}\sum_{l=1}^{m} h_W\big((X_i,R_i^X),(X_j,R_j^X);(Y_k,R_k^Y),(Y_l,R_l^Y)\big), \tag{6.13}$$

where

$$h_W\big((X_1,R_1^X),(X_2,R_2^X);(Y_1,R_1^Y),(Y_2,R_2^Y)\big)=\rho_W\big((X_1,R_1^X),(Y_2,R_2^Y)\big)+\rho_W\big((X_2,R_2^X),(Y_1,R_1^Y)\big) \\ -\rho_W\big((X_1,R_1^X),(X_2,R_2^X)\big)-\rho_W\big((Y_1,R_1^Y),(Y_2,R_2^Y)\big).$$

T_{nm}^W is itself a V-statistic, so we may proceed further in a standard manner by deriving its asymptotic distribution. We state our conclusions in the following theorem.

THEOREM 6.2 [ALEKSIĆ, MILOŠEVIĆ (2025)]**.** *Let $X_1,\dots,X_n$ and $Y_1,\dots,Y_m$ be two independent samples of IID d-variate random vectors, and let T_{nm}^W be as in* (6.13)*. Let the response indicators R^X and R^Y and their realizations r^X and r^Y be defined as in Section 6.2. Define the integral operator $A: L^2\big(\mathbb{R}^{2d},\mathrm{d}F'(x,r^X)\big)\to L^2\big(\mathbb{R}^{2d},\mathrm{d}F'(x,r^X)\big)$, where F' is a CDF of (X,R^X) and of (Y,R^Y), as*

$$Ag\big((x_1,r_1^X),(y_1,r_1^Y)\big)=\int_{\mathbb{R}^{2d}\times\mathbb{R}^{2d}} h_W\big((x_1,r_1^X),(x_2,r_2^X);(y_1,r_1^Y),(y_2,r_2^Y)\big)\mathrm{d}F'(x_2,r_2^X)\mathrm{d}F'(y_2,r_2^Y). \tag{6.14}$$

Let $\{\lambda_j, j\geqslant 1\}$ be the sequence of eigenvalues of A and let $\{f_j, j\geqslant 1\}$ be the sequence of corresponding orthonormal eigenfunctions. Let

$$c_j=\mathbb{E}\big[f_j\big((X_1,R_1^X),(Y_1,R_1^Y)\big)f_j\big((X_1,R_1^X),(Y_2,R_2^Y)\big)\big],$$

and

$$\eta=\mathbb{E}\big[\rho_W\big((X_1,R_1^X),(Y_1,R_1^Y)\big)\big].$$

If $(X_1,R_1^X)\overset{D}{=}(Y_1,R_1^Y)$, then

$$\frac{nm}{n+m}T_{nm}^W\xrightarrow{D}\eta+2\sum_{j=1}^{\infty}\lambda_j c_j\big(\chi_{1,j}^2-1\big),\quad as\quad n,m\to\infty,\quad \frac{n}{n+m}\to\lambda^2\in(0,1),$$

where $\{\chi_{1,j}^2, j\geqslant 1\}$ are IID χ_1^2-distributed random variables.

PROOF. We first note that the assumptions of Theorem 6.2 are consistent, since the operator A from (6.14) is indeed known to be compact and self-adjoint, so its eigenvalues do form a diminishing sequence, and the eigenfunctions are orthonormal (see e.g. Henze, 2024, Ch. 8).

The kernel is weakly degenerate. Indeed, it holds that

$$h_{1,1}\big((x_1,r_1^X),(y_1,r_1^Y)\big)=\mathbb{E}\Big[\rho_W\big((x_1,r_1^X),(Y_2,R_2^Y)\big)+\rho_W\big((X_2,R_2^X),(y_1,r_1^Y)\big) \\ -\rho_W\big((x_1,r_1^X),(X_2,R_2^X)\big)-\rho_W\big((y_1,y_1^Y),(Y_2,R_2^Y)\big)\Big]=0,$$

where two pairs of terms cancel out due to symmetry of ρ_W and the fact that $(X,R^X)\overset{D}{=}(Y,R^Y)$.

The key idea of the proof is to use the results of Neuhaus (1977) that we have restated in our notation in Section 1.4. Similarly to the Lemma 2 of Fernández et al. (2008), it holds that, for every $j \geqslant 1$, $f_j\big((x, r^X),(y, r^Y)\big) = -f_j\big((y, r^Y),(x, r^X)\big)$, and, as a consequence,

$$\mathbb{E}\big[f_j\big((X_1, R_1^X),(Y_1, R_1^Y)\big) f_j\big((X_1, R_1^X),(Y_2, R_2^Y)\big)\big] = \mathbb{E}\big[f_j\big((X_1, R_1^X),(Y_1, R_1^Y)\big) f_j\big((X_2, R_2^X),(Y_1, R_1^Y)\big)\big].$$

Denote this quantity as c_j. If a_j^2 and b_j^2 are defined as in (1.10), it is readily seen that

$$\begin{aligned} a_j^2 &= (1-\lambda^2)\mathbb{E}\big[f_j\big((X_1, R_1^X),(Y_1, R_1^Y)\big) f_j\big((X_1, R_1^X),(Y_2, R_2^Y)\big)\big] = (1-\lambda^2)c_j, \\ b_j &= \lambda^2 \mathbb{E}\big[f_j\big((X_1, R_1^X),(Y_1, R_1^Y)\big) f_j\big((X_2, R_2^X),(Y_1, R_1^Y)\big)\big] = \lambda^2 c_j, \end{aligned}$$

for every $j \geqslant 1$.

If $\frac{nm}{n+m} T_{nm}^W$ is understood in the context of (1.8), for $t_1 = t_2 = 1$, then the corresponding $U(1,1)$ can be written as

$$\begin{aligned} U(1,1) &= \sum_{j=1}^{\infty} \lambda_j \Big[\big(a_j W_{1j}(1) + b_j W_{2j}(1)\big)^2 - \big(a_j^2 + b_j^2\big) \Big] \\ &= \sum_{j=1}^{\infty} \lambda_j c_j \Big[\big(\underbrace{W_{1j}(1) + W_{2j}(1)}_{\sim \mathcal{N}(0,2)} \big)^2 - 2 \Big] \\ &= 2 \sum_{j=1}^{\infty} \lambda_j c_j \big(\chi_{1,j}^2 - 1 \big). \end{aligned}$$

Next, observe that

$$\begin{aligned} \frac{nm}{n+m} T_{nm}^W &= \frac{1}{nm(n+m)} \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{m}\sum_{l=1}^{m} h_W\big((X_i, R_i^X),(X_j, R_j^X);(Y_k, R_k^Y),(Y_l, R_l^Y)\big) \\ &= \frac{1}{nm(n+m)} \sum_{i=1}^{n}\sum_{\substack{j=1 \\ j\neq i}}^{n}\sum_{k=1}^{m}\sum_{l=1}^{m} h_W\big((X_i, R_i^X),(X_j, R_j^X);(Y_k, R_k^Y),(Y_l, R_l^Y)\big) \\ &\qquad + \frac{1}{nm(n+m)} \sum_{i=1}^{n}\sum_{k=1}^{m}\sum_{l=1}^{m} h_W\big((X_i, R_i^X),(X_i, R_i^X);(Y_k, R_k^Y),(Y_l, R_l^Y)\big) \\ &= \frac{1}{nm(n+m)} \sum_{i=1}^{n}\sum_{\substack{j=1 \\ j\neq i}}^{n}\sum_{k=1}^{m}\sum_{\substack{l=1 \\ l\neq k}}^{m} h_W\big((X_i, R_i^X),(X_j, R_j^X);(Y_k, R_k^Y),(Y_l, R_l^Y)\big) \\ &\qquad + \frac{1}{nm(n+m)} \sum_{i=1}^{n}\sum_{\substack{j=1 \\ j\neq i}}^{n}\sum_{k=1}^{m} h_W\big((X_i, R_i^X),(X_j, R_j^X);(Y_k, R_k^Y),(Y_k, R_k^Y)\big) \\ &\qquad + \frac{1}{nm(n+m)} \sum_{i=1}^{n}\sum_{k=1}^{m}\sum_{l=1}^{m} h_W\big((X_i, R_i^X),(X_i, R_i^X);(Y_k, R_k^Y),(Y_l, R_l^Y)\big) \\ &= \frac{1}{nm(n+m)} \sum_{i=1}^{n}\sum_{\substack{j=1 \\ j\neq i}}^{n}\sum_{k=1}^{m}\sum_{\substack{l=1 \\ l\neq k}}^{m} h_W\big((X_i, R_i^X),(X_j, R_j^X);(Y_k, R_k^Y),(Y_l, R_l^Y)\big) \\ &\qquad + \frac{1}{nm(n+m)} \sum_{i=1}^{n}\sum_{\substack{j=1 \\ j\neq i}}^{n}\sum_{k=1}^{m} h_W\big((X_i, R_i^X),(X_j, R_j^X);(Y_k, R_k^Y),(Y_k, R_k^Y)\big) \\ &\qquad + \frac{1}{nm(n+m)} \sum_{i=1}^{n}\sum_{k=1}^{m}\sum_{\substack{l=1 \\ l\neq k}}^{m} h_W\big((X_i, R_i^X),(X_i, R_i^X);(Y_k, R_k^Y),(Y_l, R_l^Y)\big) \end{aligned}$$

$$
\begin{aligned}
&\quad + \frac{1}{nm(n+m)} \sum_{i=1}^{n}\sum_{k=1}^{m} h_W\big((X_i,R_i^X),(X_i,R_i^X);(Y_k,R_k^Y),(Y_k,R_k^Y)\big) \\
&= \frac{1}{nm(n+m)} \sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n}\sum_{k=1}^{m}\sum_{\substack{l=1\\ l\neq k}}^{m} h_W\big((X_i,R_i^X),(X_j,R_j^X);(Y_k,R_k^Y),(Y_l,R_l^Y)\big) \\
&\quad + \left(\frac{n}{n+m}-\frac{1}{n+m}\right)\frac{1}{n(n-1)m} \sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n}\sum_{k=1}^{m} h_W\big((X_i,R_i^X),(X_j,R_j^X);(Y_k,R_k^Y),(Y_k,R_k^Y)\big) \\
&\quad + \left(\frac{m}{n+m}-\frac{1}{n+m}\right)\frac{1}{nm(m-1)} \sum_{i=1}^{n}\sum_{k=1}^{m}\sum_{\substack{l=1\\ l\neq k}}^{m} h_W\big((X_i,R_i^X),(X_i,R_i^X);(Y_k,R_k^Y),(Y_l,R_l^Y)\big) \\
&\quad + \frac{1}{n+m}\frac{1}{nm} \sum_{i=1}^{n}\sum_{k=1}^{m} h_W\big((X_i,R_i^X),(X_i,R_i^X);(Y_k,R_k^Y),(Y_k,R_k^Y)\big) \\
&= M_{nm} + \left(\frac{n}{n+m}-\frac{1}{n+m}\right)N_{nm} + \left(\frac{m}{n+m}-\frac{1}{n+m}\right)P_{nm} + \frac{1}{n+m}Q_{nm},
\end{aligned}
$$

where

$$
M_{nm} = \frac{1}{nm(n+m)} \sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n}\sum_{k=1}^{m}\sum_{\substack{l=1\\ l\neq k}}^{m} h_W\big((X_i,R_i^X),(X_j,R_j^X);(Y_k,R_k^Y),(Y_l,R_l^Y)\big),
$$

$$
N_{nm} = \frac{1}{n(n-1)m} \sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n}\sum_{k=1}^{m} h_W\big((X_i,R_i^X),(X_j,R_j^X);(Y_k,R_k^Y),(Y_k,R_k^Y)\big),
$$

$$
P_{nm} = \frac{1}{nm(m-1)} \sum_{i=1}^{n}\sum_{k=1}^{m}\sum_{\substack{l=1\\ l\neq k}}^{m} h_W\big((X_i,R_i^X),(X_i,R_i^X);(Y_k,R_k^Y),(Y_l,R_l^Y)\big),
$$

and

$$
Q_{nm} = \frac{1}{nm} \sum_{i=1}^{n}\sum_{k=1}^{m} h_W\big((X_i,R_i^X),(X_i,R_i^X);(Y_k,R_k^Y),(Y_k,R_k^Y)\big).
$$

It is clear that, as $n,m\to\infty$, $n/(n+m)\to\lambda^2$,

$$
\left(\frac{n}{n+m}-\frac{1}{n+m}\right)\to\lambda^2, \quad \left(\frac{m}{n+m}-\frac{1}{n+m}\right)\to 1-\lambda^2, \quad \frac{1}{n+m}\to 0. \tag{6.15}
$$

Next, by (1.11), we have that

$$
M_{nm} \xrightarrow{D} U(1,1). \tag{6.16}
$$

By the Law of Large Numbers for U-statistics, we have that

$$
N_{nm} \xrightarrow{P} \mathbb{E}\big(h_W\big((X_1,R_1^X),(X_2,R_2^X);(Y_1,R_1^Y),(Y_1,R_1^Y)\big)\big), \tag{6.17}
$$

$$
P_{nm} \xrightarrow{P} \mathbb{E}\big(h_W\big((X_1,R_1^X),(X_1,R_1^X);(Y_1,R_1^Y),(Y_2,R_2^Y)\big)\big), \tag{6.18}
$$

and

$$Q_{nm} \xrightarrow{P} \mathbb{E}\left(h_W\left((X_1, R_1^X), (X_1, R_1^X); (Y_1, R_1^Y), (Y_1, R_1^Y)\right)\right). \tag{6.19}$$

Combining (6.15), (6.16), (6.17), (6.18), and (6.19), we conclude that

$$\begin{aligned}\frac{nm}{n+m} T_{nm}^W &\xrightarrow{D} U(1,1) + \lambda^2 \mathbb{E}\left(h_W\left((X_1, R_1^X), (X_2, R_2^X); (Y_1, R_1^Y), (Y_1, R_1^Y)\right)\right) \\ &\quad + (1-\lambda^2)\mathbb{E}\left(h_W\left((X_1, R_1^X), (X_1, R_1^X); (Y_1, R_1^Y), (Y_2, R_2^Y)\right)\right) \\ &= U(1,1) + \mathbb{E}\left(h_W\left((X_1, R_1^X), (X_2, R_2^X); (Y_1, R_1^Y), (Y_1, R_1^Y)\right)\right),\end{aligned}$$

where the last equality is due to the symmetry of h_W and due to the assumption that $(X, R_X) \stackrel{D}{=} (Y, R_Y)$. Noting that

$$\mathbb{E}\left(h_W\left((X_1, R_1^X), (X_2, R_2^X); (Y_1, R_1^Y), (Y_1, R_1^Y)\right)\right) = \mathbb{E}\left[\rho_W\left((X_1, R_1^X), (Y_1, R_1^Y)\right)\right]$$

concludes the proof of the theorem. ■

Remark 6.1. Note that Theorem 6.2 does not assume the MCAR assumption. However, it is clear that, if it holds, then it is sufficient to assume that $R^X \stackrel{D}{=} R^Y$ and that the null hypothesis holds.

The null asymptotic distribution of the test statistic T_{nm}^W clearly depends on the distribution of (X_1, R_1^X). However, it is important to note that the use of random weights did not affect the type of the limiting distribution.

Given the above, turning to resampling procedures is a straightforward decision in this case as well. Those procedures will be presented in Subsection 6.2.4.

6.2.3 Imputation

In many practical situations, a particular dataset will not be used exclusively for a single statistical procedure, such as two-sample testing, but rather as a subject of a broader range of analyses. This makes imputation particularly appealing: by filling in the missing values and producing a completed dataset, it allows analysts to apply standard methods, possibly known not to be sensitive to the fact that the data are imputed, without needing to account for missingness at each step. Moreover, in real-world scenarios, people that analyze data may not have specialized knowledge of missing data techniques or access to tools that handle incomplete observations correctly. So the only option for them would be to treat the imputed dataset as complete. For this reason, it is often desirable to provide a single imputed version of the dataset that can be used in lot of the future analyses, including hypothesis testing and many others. An algorithm that imputes the data is a natural requirement in such scenarios. In a similar manner as for the statistic T_{nm}^W, we turn to the bootstrap once again. One such algorithm is proposed in the Subsection 6.2.4.

6.2.4 Resampling procedures

We propose two bootstrap resampling procedures that can be used to both T_{nm}^{CC} and T_{nm}^W, resulting in four distinct two-sample testing procedures overall. The first bootstrap approach, summarized in Algorithm 6.1 (replacing the generic T with T_{nm}^{CC} or T_{nm}^W), is designed to account for the structure of the incomplete data by treating complete and incomplete cases separately during resampling. Specifically, the pooled dataset is divided into two subsets: one containing only complete cases and the other containing only incomplete ones. Resampling is then

carried out independently within each subset, after which the resampled complete and incomplete cases are recombined to form new bootstrap samples. This procedure ensures that the proportion of complete cases in each bootstrap sample remains close to the one in the original data. By preserving this proportion, the algorithm keeps the original missingness pattern and avoids artificially inflating or deflating the amount of information available in the bootstrap replicates, compared to that in the original incomplete sample.

In addition to the first method, we also propose a simpler alternative, described in Algorithm 6.2. Unlike the previous approach, this algorithm does not distinguish between complete and incomplete cases during resampling. Instead, it pools all available observations and randomly splits them into two bootstrap samples of fixed sizes n and m. This method significantly reduces the complexity of the resampling procedure by avoiding the need to track and preserve the proportion of complete cases. It also slightly accelerates computation. Although this is often unimportant in applied settings, it becomes critical in simulation studies, where the test must be rerun tens or hundreds of thousands of times.

However, this simplicity may come at a cost. Since the proportion of complete to incomplete cases is not preserved across bootstrap samples, the testing procedures may have difficulties controlling the type I error, or may exhibit reduced power in certain settings. One of the objectives of the simulation study that follows will be to examine whether this potential trade-off between computational simplicity and statistical performance has an impact in practice. In particular, we aim to assess whether the test remains well-calibrated and retains sufficient power under various missingness scenarios.

Algorithm 6.1 A bootstrap algorithm for the energy test: preserving the proportions of complete cases.

1: Start with incomplete samples $x = (x_1, \ldots, x_n)$ and $y = (y_1, \ldots, y_m)$ of d-variate vectors;
2: Calculate the value $T(x, y)$ of the test statistic T;
3: Produce two pooled samples: z_{com} that consists of complete cases from both x and y, and z_{inc} that consists of incomplete cases;
4: Randomly split z_{com} into x^*_{com} of size $\hat{n}$ and y^*_{com} of size $\hat{m}$; randomly split z_{inc} into x^*_{inc} of size $n-\hat{n}$ and y^*_{inc} of size $m-\hat{m}$;
5: Combine x^*_{com} and x^*_{inc} into x^*; combine y^*_{com} and y^*_{inc} into y^*;
6: Calculate $T^* = T(x^*, y^*)$;
7: Repeat the steps 4-6 B times to obtain $T^*_1, T^*_2, \ldots, T^*_B$;
8: Reject the null hypothesis at the significance level α if $T(x, y)$ is greater than the $(1-\alpha)$-quantile of the empirical bootstrap distribution of $(T^*_1, T^*_2, \ldots, T^*_B)$.

Algorithm 6.2 A bootstrap algorithm for the energy test: resampling directly from the pooled sample.

1: Start with incomplete samples $x = (x_1, \ldots, x_n)$ and $y = (y_1, \ldots, y_m)$ of d-variate vectors;
2: Calculate the value $T(x, y)$ of the test statistic;
3: Combine incomplete samples x and y to obtain the pooled sample $z = (x, y)$;
4: Randomly split the pooled sample z into x^* of size n, and y^* of size m;
5: Calculate the value $T^* = T(x^*, y^*)$;
6: Repeat the steps 4-5 B times to obtain $T^*_1, T^*_2, \ldots, T^*_B$;
7: Reject the null hypothesis at the significance level α if $T(x, y)$ is greater than the $(1-\alpha)$-quantile of the empirical bootstrap distribution of $(T^*_1, T^*_2, \ldots, T^*_B)$.

Under the imputation approach, we introduce the Algorithm 6.3. The algorithm begins by imputing the original incomplete samples resulting in their fully observed versions, from

which the test statistic is computed. To approximate its null distribution via bootstrap, the algorithm pools the original data and resamples from it to create new incomplete bootstrap samples, which are then imputed using the same method method of imputation. The test statistic is calculated on each imputed bootstrap pair, and the distribution of these replicates is used to determine the critical value. This approach allows the test to operate on completed data while remaining coherent with the imputation model throughout the resampling process.

The success of this approach, however, depends critically on the quality of the imputation: poor or biased imputations may distort the type I error or reduce the power, as we will see from the results of our simulations. Therefore, one of the goals of the simulation study will be to examine the sensitivity of this method to the choice of imputation strategy and to compare its performance with the other two algorithms. This will help clarify the trade-offs involved in choosing a more general-purpose imputation approach over more specific, tailored, weighting methods.

Algorithm 6.3 A bootstrap algorithm for the energy test: imputing the data.

1: Start with incomplete samples $x = (x_1, \ldots, x_n)$ and $y = (y_1, \ldots, y_m)$ of d-variate vectors;
2: Impute the samples using the chosen method to obtain x_{imp} and y_{imp};
3: Calculate the value $T_{nm}^{IMP}(x_{imp}, y_{imp})$ of the test statistic T_{nm} from (6.4);
4: Combine incomplete samples x and y to obtain the pooled sample $z = (x, y)$;
5: Randomly split pooled sample z into x^* of size n, and y^* of size m;
6: Impute x^* and y^* using the chosen method to obtain x^*_{imp} and y^*_{imp};
7: Calculate the value $T_{nm}^{IMP,*} = T_{nm}(x^*_{imp}, y^*_{imp})$;
8: Repeat the steps 5-7 B times to obtain $T_{nm,1}^{IMP,*}, T_{nm,2}^{IMP,*}, \ldots, T_{nm,B}^{IMP,*}$;
9: Reject the null hypothesis at the significance level α if $T_{nm}^{IMP}(x_{imp}, y_{imp})$ is greater than the $(1-\alpha)$-quantile of the empirical bootstrap distribution of $(T_{nm,1}^{IMP,*}, T_{nm,2}^{IMP,*}, \ldots, T_{nm,B}^{IMP,*})$.

6.3 Empirical study

In this section, we present the results of an empirical study conducted under the MCAR setting, designed to evaluate the performance of the proposed methods. Simulated data were used to compare testing procedures under controlled scenarios and limited computational resources. We note that the scenarios presented here represent only a subset of those studied. A more extensive investigation, including various MAR settings, as well as analysis on real data, is available in the accompanying paper by Aleksić and Milošević (2025b). These additional results are omitted from the thesis to avoid overloading the main text, as they offer limited added value in terms of methodological novelty.

6.3.1 Design of the study

Due to the high computational demands of these methods, designing the study was a non-trivial task: it needed to cover as many scenarios as possible while minimizing the number of cases that need to be simulated. One of the first trade-offs that had to be made was the use of the warp-speed bootstrap algorithm (Giacomini et al., 2013) in place of the classical bootstrap method, in order to avoid nested loops during the Monte Carlo simulations. For the convenience of the reader, we restate the general warp-speed bootstrap procedure in Algorithm 6.4. Given that $N = 5000$ replicates were used, this approximation is not expected to have a substantial impact on the simulation results. Regarding the results presented in this text, the sample sizes were fixed to be $n = 100$ and $m = 50$, and trivariate data were considered.

Algorithm 6.4 Warp-speed bootstrap algorithm for approximating type I error or power.

1: Generate samples $x = (x_1, \ldots, x_n)$ of size n and $y = (y_1, \ldots, y_m)$ of size m from the assumed distributions;
2: Calculate the test statistic $\hat{T}$ from x and y;
3: Generate one single bootstrap resample x^* from and one y^* as in one of the Algorithms from Subsection 6.2.4;
4: Calculate the bootstrap test statistic $\hat{T}^*$ from x^* and y^*;
5: Repeat steps 1–4 N times to obtain pairs $(\hat{T}_1, \hat{T}_1^*), \ldots, (\hat{T}_N, \hat{T}_N^*)$;
6: Calculate the estimated type I error (or power) of the bootstrap test at significance level α as the average rejection rate:

$$\widehat{\text{rate}} = \frac{1}{N} \sum_{k=1}^{N} I\left\{T_k \geqslant q^*\right\},$$

where q^* is the empirical $(1-\alpha)$-quantile of the distribution of $T_1^*, \ldots, T_N^*$, and α is the nominal level.

As shown in Table 6.1, which serves as a legend for the other tables, combining Algorithms 6.1 and 6.2 with both T_{nm}^{CC} and T_{nm}^{W} yields four testing procedures. In addition, Algorithm 6.3 is paired with commonly used imputation methods: mean and median imputation (from `R` package `missMethods` by Rockel, 2023), kNN imputation (from `R` package `bnstruct` by Franzin et al., 2017), and missForest (from `R` package `missForest` by Stekhoven, 2013). This leaves us with a total of eight testing procedures to be evaluated in this study.

As stated previously, this thesis focuses on the MCAR mechanism. However, results for three MAR mechanisms are available in the accompanying paper (Aleksić and Milošević, 2025b). The first two mechanisms, *MAR 1 to 9* and *MAR rank*, are implemented in the `missMethods` `R` package (Rockel, 2023) and have been used in recent studies (e.g., Bordino and Berrett, 2024; Aleksić, 2024, 2025a). For detailed explanations of these mechanisms, we refer to Santos et al. (2019), and for a general overview of generating missing data we refer to the monograph by Van Buuren (2018). The third, *MAR logistic*, was implemented from scratch, and it assigns missingness based on a logistic regression model, with control variables as predictors. We summarize the findings for these settings in Remark 6.2.

The results for the trivariate data will be presented here, while the results for the decavariate case can be found in the aforementioned paper.

Certainly, there are infinitely many ways in which the distributions of X and Y can differ. However, two of the most commonly studied types of differences are shifts in the mean and changes in variance. The matrices

$$C_1 = \begin{bmatrix} 0.5 & 0 & 0 \\ 0 & 0.5 & 0 \\ 0 & 0 & 0.5 \end{bmatrix} \quad \text{and} \quad C_2 = \begin{bmatrix} 1 & 0.5 & 0.5 \\ 0.5 & 1 & 0.5 \\ 0.5 & 0.5 & 1 \end{bmatrix}$$

were used as covariance matrices, together with the standard identity matrix. Besides the standard zero mean, we considered the mean vector $m_1 = (0.5, 0.5, 0.5)$ to see how the testing procedures detect the change in mean.

To assess both light- and heavy-tailed distributions in different dimensions, we considered five non-degenerate normal distributions and Student's t distribution with 5 degrees of freedom. Specifically, the distributions are: $\mathcal{N}_3(0, I)$, $\mathcal{N}_3(0, C_1)$, $\mathcal{N}_3(m_1, C_1)$, $\mathcal{N}_3(0, C_2)$, $\mathcal{N}_3(m_1, C_2)$, and $t_5(0, I)$, where I is the identity matrix of size 3×3.

We considered three missingness settings: two with equal missingness probabilities across all three variables (0.1 and 0.4), and one with unequal probabilities of 0.1, 0.2, and 0.3.

6.3.2 Results

With the structure of the study now thoroughly outlined, we are in a position to present and interpret the results obtained from the conducted simulations.

Table 6.1: The legend of simulation procedures.

%	Distribution of Y	
Distribution of X	Algorithm 6.1 with T_{nm}^{CC}	Algorithm 6.3 with mean imputation
	Algorithm 6.2 with T_{nm}^{CC}	Algorithm 6.3 with median imputation
	Algorithm 6.1 with T_{nm}^{W}	Algorithm 6.3 with 6NN imputation
	Algorithm 6.2 with T_{nm}^{W}	Algorithm 6.3 with missForest imputation

Table 6.2 presents the empirical type I errors and powers under the MCAR mechanism, with a missingness probability of 0.1 for each variable. All eight testing procedures are generally well calibrated. The procedures based on the weighted test statistic show a slightly elevated type I error in some cases, but the deviation is not major and does not undermine the overall validity of the procedures. Interestingly, when the missingness probability is increased to 0.4 (Table 6.3), the type I error does not increase; in fact, it appears slightly improved. In the setting with unequal missingness probabilities across variables (0.1, 0.2, and 0.3), shown in Table 6.4, all of the studied methods remain acceptably calibrated.

Table 6.2: Percentage of rejections (rounded to the nearest integer) for trivariate data missing according to the MCAR mechanism, $p = (0.1, 0.1, 0.1)$, $n = 100$, $m = 50$, $N = 5000$.

%	$\mathcal{N}_3(0, I)$		$\mathcal{N}_3(0, C_1)$		$\mathcal{N}_3(m_1, C_1)$		$\mathcal{N}_3(0, C_2)$		$\mathcal{N}_3(m_1, C_2)$		$t_5(0, I)$	
$\mathcal{N}_3(0, I)$	4	4	25	25	100	100	13	11	92	96	12	9
	4	4	24	24	100	100	11	11	91	96	12	9
	5	5	40	14	100	97	14	8	96	81	13	9
	5	4	38	26	100	100	14	15	96	96	14	10
$\mathcal{N}_3(0, C_1)$	47	46	5	5	100	100	53	53	99	100	77	80
	45	48	5	5	100	100	54	53	98	100	77	80
	55	34	5	4	100	100	61	38	99	96	76	59
	56	50	4	5	100	100	62	65	100	100	77	83
$\mathcal{N}_3(m_1, C_1)$	100	100	100	100	5	5	99	99	54	55	100	100
	100	100	100	100	5	4	99	99	57	55	100	100
	100	99	100	99	5	5	99	96	58	37	100	100
	100	100	100	100	5	5	99	100	60	67	100	100
$\mathcal{N}_3(0, C_2)$	11	9	31	35	99	100	5	6	80	89	21	18
	10	9	30	34	99	100	5	6	80	89	19	17
	13	6	44	14	100	93	6	5	88	68	25	13
	13	13	47	51	100	100	6	6	88	89	35	24
$\mathcal{N}_3(m_1, C_2)$	94	97	99	100	34	28	80	89	5	5	94	97
	93	97	99	100	32	27	80	89	5	6	94	97
	97	82	100	94	45	13	88	68	5	5	97	82
	97	98	100	100	45	46	88	89	5	5	96	97
$t_5(0, I)$	6	5	57	56	100	100	12	12	90	95	5	5
	5	6	50	58	99	100	11	12	90	95	5	6
	7	5	65	22	99	98	18	8	93	74	5	5
	7	6	63	58	99	100	17	15	94	95	5	6

Analyzing the empirical power, kNN imputation can immediately be ruled out, having the empirical power substantially lower than any other. Procedures that use weighted test statistic performed the best overall. Mean and median imputation follow closely in some settings, but are mostly lacking power compared to the former two. Imputation using the missForest algorithm performs somewhere in between. A similar pattern is observed for the settings with unequal missingness probabilities (Table 6.4).

In Subsection 6.2.4, it was noted that Algorithm 6.2 represents a simplified version of Algorithm 6.1, and that this simplification could potentially come at a cost in terms of accuracy or validity. However, as evidenced by the results presented in Tables 6.2 and 6.3, the performance of both algorithms is essentially comparable.

Table 6.3: Percentage of rejections (rounded to the nearest integer) for trivariate data missing according to the MCAR mechanism, $p = (0.4, 0.4, 0.4)$, $n = 100$, $m = 50$, $N = 5000$.

%	$\mathcal{N}_3(0, I)$		$\mathcal{N}_3(0, C_1)$		$\mathcal{N}_3(m_1, C_1)$		$\mathcal{N}_3(0, C_2)$		$\mathcal{N}_3(m_1, C_2)$		$t_5(0, I)$	
$\mathcal{N}_3(0, I)$	5	5	4	6	5	98	6	6	35	86	9	9
	5	5	4	6	5	95	6	5	33	80	9	7
	5	4	22	3	10	23	8	5	79	20	11	8
	5	5	20	7	10	94	7	8	79	78	10	9
$\mathcal{N}_3(0, C_1)$	14	15	5	5	63	100	15	14	54	95	24	26
	14	16	5	6	65	99	15	16	54	93	24	26
	33	12	5	5	99	42	35	12	93	37	52	19
	32	19	5	5	99	98	36	30	93	95	52	36
$\mathcal{N}_3(m_1, C_1)$	60	98	65	100	4	5	54	95	17	15	65	98
	63	96	67	99	5	5	52	92	15	15	65	97
	96	39	98	39	5	5	93	37	36	11	95	39
	96	96	99	98	5	5	93	95	36	32	85	97
$\mathcal{N}_3(0, C_2)$	6	5	4	5	42	96	6	5	3	80	10	8
	6	5	4	5	40	92	5	5	28	75	10	7
	7	4	22	4	93	23	5	5	72	21	14	9
	7	10	22	19	93	95	5	5	79	71	15	1
$\mathcal{N}_3(m_1, C_2)$	35	86	40	95	5	5	27	80	5	5	34	82
	35	79	40	91	5	6	27	73	6	4	34	78
	81	22	93	22	22	3	71	20	5	5	79	21
	81	82	93	94	22	19	71	70	5	5	78	82
$t_5(0, I)$	3	4	5	6	39	95	4	4	28	78	5	5
	4	5	5	8	39	95	5	5	27	77	5	5
	5	3	35	4	92	12	9	4	72	14	4	5
	5	5	33	10	92	92	9	6	73	12	4	5

Taking all of this into consideration, under the MCAR missingness mechanism, we recommend employing Algorithm 6.2 in combination with the weighted test statistic T^W_{nm}. This combination consistently delivers the best overall performance with respect to empirical type I error control and statistical power, balancing simplicity and effectiveness.

Remark 6.2. As mentioned earlier, beyond the MCAR setting studied in this thesis, more complex MAR mechanisms were examined in the accompanying paper (Aleksić and Milošević, 2025b). In general, weighting-based testing procedures tend to show the best power when the missingness rate is low; they may suffer from inflated type I error as missingness increases, but only under certain missingness mechanisms. Imputation methods, particularly mean and median imputation, maintain better control of type I error at higher missingness rates, though their power can be inconsistent: sometimes slightly outperforming weighting methods, but more often falling substantially behind. The missForest imputation typically achieves a good balance between power and type I error control, performing substantially better than nearest-neighbor imputation. Although nearest-neighbor imputation preserves nominal type I error rates, it is consistently outperformed in terms of power and should therefore be avoided.

Our preliminary simulations show that none of the methods had satisfactory type I error control under the MNAR upper censoring setting, and adapting the energy test statistic for such scenarios remains an open problem.

Table 6.4: Percentage of rejections (rounded to the nearest integer) for trivariate data missing according to the MCAR mechanism, $p = (0.1, 0.2, 0.3)$, $n = 100$, $m = 50$, $N = 5000$.

%	$\mathcal{N}_3(0, I)$		$\mathcal{N}_3(0, C_1)$		$\mathcal{N}_3(m_1, C_1)$		$\mathcal{N}_3(0, C_2)$		$\mathcal{N}_3(m_1, C_2)$		$t_5(0, I)$	
$\mathcal{N}_3(0, I)$	5	5	12	15	96	100	9	7	78	93	8	8
	5	5	12	14	95	100	9	7	77	91	9	7
	5	5	31	6	99	53	11	6	92	39	11	8
	5	4	32	17	99	100	11	11	92	94	11	9
$\mathcal{N}_3(0, C_1)$	30	31	5	5	99	100	35	31	92	99	59	60
	31	31	5	5	99	100	34	31	93	99	57	59
	48	14	5	5	100	73	52	16	99	63	71	24
	44	37	4	5	100	100	49	55	99	99	69	68
$\mathcal{N}_3(m_1, C_1)$	98	100	99	100	5	5	92	99	36	24	98	100
	97	100	99	100	5	6	92	98	36	31	98	100
	99	70	100	73	5	5	98	64	51	15	99	72
	99	100	100	100	5	5	98	99	51	54	100	100
$\mathcal{N}_3(0, C_2)$	9	6	17	16	91	99	5	5	64	88	16	12
	8	7	15	14	91	99	5	5	64	86	16	12
	10	5	38	5	98	46	4	5	85	36	22	7
	10	11	37	39	98	100	5	5	86	88	20	21
$\mathcal{N}_3(m_1, C_2)$	78	95	92	99	16	14	64	88	6	6	80	94
	79	93	93	99	16	14	64	85	6	5	79	93
	94	40	98	44	37	6	85	36	5	5	93	36
	94	95	98	100	37	37	84	86	5	6	92	95
$t_5(0, I)$	4	5	27	29	97	100	8	7	73	91	5	6
	5	6	27	30	97	100	7	8	73	90	6	6
	8	5	55	6	98	37	14	4	89	25	6	5
	7	5	55	37	98	100	16	11	88	92	5	6

Chapter 7

Conclusions and future work

This thesis is devoted to the problem of model specification testing when the data contain missing values. To make the dissertation self-contained, Chapters 1 and 2 introduced essential mathematical and statistical background on U-statistics, V-statistics and missing data analysis, which are critical for understanding the methods and results presented in the subsequent Chapters 3–6. Building upon this foundational material, what follows is a summary of the main contributions, findings, and possible extensions for each core chapter of this thesis.

In Chapter 3, we introduced two novel statistical tests for assessing the MCAR assumption: the second one being the generalized version of the first, with its own merits and flaws. Across the majority of the scenarios examined, particularly those that are more likely to arise in practice, such as cases with moderate missingness rates and a large number of variables, the proposed tests consistently outperformed Little's MCAR test. In these settings, it demonstrated superior control of the type I error rate, higher statistical power, and greater robustness to violations of the assumption of finite fourth moments, provided that both tests performed satisfactorily.

In situations involving infinite fourth moments combined with the alternatives that are more difficult to detect, both the novel test and Little's test exhibited unexpected behavior. Specifically, their power declined as the missingness rate increased, which was unexpected.

However, in contrast to Little's test, the novel test did not exhibit a loss of power as dimensionality increased, indicating that it performs more reliably in high-dimensional settings. This stability suggests that the novel test may be better suited for modern applications involving large number of variables relative to the sample size, where traditional methods often struggle.

A natural direction for future research would be to investigate the asymptotic properties of the two proposed tests as well as Little's test as the dimension grows. In particular, it would be of interest to derive their asymptotic distributions when the dimension tends to infinity: either at the same rate as, faster than, or slower than the sample size.

With regard to Remark 3.9, another possible direction for further improvement would be to replace the covariance, which is used in the current formulation as a measure of linear dependence, with an alternative discrepancy measure that either characterizes dependence or is more closely related to it. This is a potential goal for future research.

Chapter 4 was devoted to the study of Kendall's independence test in the presence of MCAR data. The contributions of that chapter can be summarized in three main aspects. First, we derived the limiting distribution of U-statistics with a non-degenerate kernel of order two under the MCAR setting, and we applied these results to the well-known Kendall's test statistic for testing independence.

Second, we established the limiting distribution of Kendall's test statistic when the widely used median-based imputation method is applied to handle missing values.

Third, we carried out a comparative analysis of two approaches for handling missing data,the complete-case approach and the median-based imputation approach, in the context of Kendall's tau. Their performance was evaluated in terms of empirical type I error and statistical power, which are the most relevant criteria for assessing the practical effectiveness of statistical tests.

In summary, our results showed that the median-based approach performed more reliably for smaller samples, making it a sensible choice in such situations. One possible drawback of this approach is that it could slightly distort the estimation of Kendall's tau, which might then affect the power of the test. In our study, however, this effect was small. For larger samples, the complete-case approach gave slightly better results. Finally, our simulations confirmed that the way missingness was introduced in the data had a clear effect on how well the test performed.

Exploring how advanced imputation methods affect the statistical properties of Kendall's test of independence would be a promising direction for future research. It would also be valuable to study the asymptotic properties of various degenerate U- and V-statistics that are commonly used in various areas of model specification testing.

Chapter 5 focused on testing MVN in the presence of MCAR data using the BHEP test. The contributions of this chapter can be summarized as follows. First, we proved that complete-case analysis can be applied for MVN testing under MCAR data, since in this case the test statistic has the same asymptotic distribution as the test statistic computed on a fully observed sample. Second, we examined the limiting distribution of the test statistic when imputation is used and showed that, under such procedures, the affine invariance property of the statistic is no longer preserved. However, we also noted that, for carefully chosen parameter estimators, the distribution may remain independent of the unknown parameters.

To address this loss of invariance, we proposed a bootstrap algorithm for MVN testing that maintains proper type I error control. We also emphasized the potential problems arising from the common practice of treating an imputed dataset as complete and carrying out the analysis without accounting for the imputation process. Finally, we compared the power of the BHEP test under the complete-case approach and under several common imputation methods, including mean, median, and kNN imputation.

As demonstrated in the real-data example in Section 5.4, the complete-case approach proved effective in detecting departures from MVN. Although our power study showed that, in general, the mean and median imputation approaches achieved higher power, we recommend using the complete-case approach whenever the sample size is sufficiently large. The main reasons for this recommendation are its simplicity, interpretability, and computational efficiency. We also advise exercising caution when working with samples containing a small number of observations.

One natural extension of this work would be to investigate other commonly used multivariate tests for normality and, beyond that, to study the behavior of recent goodness-of-fit tests for other multivariate distributions (e.g. Karling et al., 2023; Ebner et al., 2024), especially for the data with dependent observations (Meintanis et al., 2024). Another direction would be to examine the properties of such tests under MAR settings. Our preliminary findings indicate that, in those settings, all of the approaches considered in Chapter 5 do not have a satisfactory performance. Developing a bootstrap algorithm that effectively addresses this scenario remains an open problem. Furthermore, preliminary simulation results, included in the Supplementary Material of Aleksić and Milošević (2025a), suggested that the null distribution of the test statistic of the BHEP test, when calculated on an imputed dataset and scaled with parameter estimates obtained from the same dataset, might not depend on the mean vector or covariance matrix of the underlying multivariate normal distribution. This observation points to another possible research direction, namely, the study of the invariance properties of different MVN tests under various imputation strategies and missingness mechanisms.

In Chapter 6, we adapted the well-known energy-based two-sample test to handle data that are not necessarily MCAR. Our results showed that energy-based two-sample tests, when properly modified, remain effective in the presence of missing data. Among the approaches we examined, the weighted method stood out due to its ability to utilize all available observations and its superior power performance. This advantage is particularly important because the weighted approach can be readily applied to other distance-based tests, thereby broadening its practical applicability.

The favorable performance of the weighted method as data dimensionality increases naturally raises the important question of how it behaves as the dimension tends to infinity. This remains a promising area for future research. Regarding imputation approaches, careful selection is crucial, since some popular methods, such as kNN, can significantly reduce the power of the test.

Heinze et al. (2024) proposed four phases of methodological research that, although developed primarily for the biostatistical framework, are broadly relevant across statistics. Their brief overview is as follows. Phase I involves the theoretical development of a new method. Phase II focuses on empirical evaluation in a narrow setting. Phase III includes validation across diverse scenarios and the creation of user-friendly software implementations of the proposed methods. Phase IV aims for comprehensive understanding of the method, including knowing when it is preferred or not, identifying common pitfalls, and developing practical diagnostics of whether the assumptions of the method are met.

Our work currently falls between Phase II and Phase III. By the end of Phase III, a user-friendly software implementation is expected, which we aim to provide in the near future. Phase IV involves gaining a deeper understanding of the method through practical use. We expect that as our methods are adopted more widely, both their strengths and limitations will become clearer.

References

Alba-Fernández, M. V., A. Batsidis, M.-D. Jiménez-Gamero, and P. Jodrá (2017). A class of tests for the two-sample problem for count data. *Journal of Computational and Applied Mathematics 318*, 220–229.

Aleksić, D. (2024). A Novel Test of Missing Completely at Random: U-statistics-based Approach. *Statistics 51*(5), 2170–2193.

Aleksić, D., M. Cuparić, and B. Milošević (2023). Non-degenerate U-statistics for data missing completely at random with application to testing independence. *Stat 12*(1), e634.

Aleksić, D. G. (2025a). A generalization of a U-statistics-based MCAR Test: Utilizing Partially Observed Variables. *arXiv preprint arXiv:2501.05596*.

Aleksić, D. G. (2025b). `GitHub` page: `https://github.com/danijel-g-aleksic`.

Aleksić, D. G. and B. Milošević (2025a). To impute or not? Testing multivariate normality on incomplete dataset: revisiting the BHEP test. *Journal of Applied Statistics 52*(9), 1742–1759.

Aleksić, D. G. and B. Milošević (2025b). Two-Sample Testing with Missing Data via Energy Distance: Weighting and Imputation Approaches. *arXiv preprint arXiv:2508.11421*.

Anderson, T. W. (1962). On the distribution of the two-sample Cramer-von Mises criterion. *The Annals of Mathematical Statistics 33*(3), 1148–1159.

Anderson, T. W. and D. A. Darling (1954). A test of goodness of fit. *Journal of the American Statistical Association 49*(268), 765–769.

Baringhaus, L. and C. Franz (2004). On a new multivariate two-sample test. *Journal of Multivariate Analysis 88*(1), 190–206.

Baringhaus, L. and N. Henze (1988). A consistent test for multivariate normality based on the empirical characteristic function. *Metrika 35*, 339–348.

Batsidis, A., N. Martín, L. Pardo, and K. Zografos (2016). ϕ-divergence based procedure for parametric change-point problems. *Methodology and Computing in Applied Probability 18*(1), 21–35.

Berrett, T. B. and R. J. Samworth (2023). Optimal nonparametric testing of Missing Completely At Random and its connections to compatibility. *The Annals of Statistics 51*(5), 2170–2193.

Billingsley, P. (1968). *Convergence of probability measures*. New York: John Wiley & Sons.

Bojinov, I. I., N. S. Pillai, and D. B. Rubin (2020). Diagnosing missing always at random in multivariate data. *Biometrika 107*(1), 246–253.

Bordino, A. and T. B. Berrett (2024). Tests of Missing Completely At Random based on sample covariance matrices. *arXiv preprint arXiv:2401.05256*.

Brislawn, C. (1988). Kernels of trace class operators. *Proceedings of the American Mathematical Society 104*(4), 1181–1190.

Brislawn, C. (1991). Traceable integral kernels on countably generated measure spaces. *Pacific Journal of Mathematics 150*(2), 229–240.

Chassan, M. and D. Concordet (2023). How to test the missing data mechanism in a hidden Markov model. *Computational Statistics & Data Analysis 182*, 107–723.

Chen, F., S. G. Meintanis, and L. Zhu (2019). On some characterizations and multidimensional criteria for testing homogeneity, symmetry and independence. *Journal of Multivariate Analysis 173*, 125–144.

Chen, H. Y. and R. Little (1999). A test of missing completely at random for generalised estimating equations with missing data. *Biometrika 86*(1), 1–13.

Chu, L. and X. Dai (2024). Manifold energy two-sample test. *Electronic Journal of Statistics 18*(1), 145–166.

Cramér, H. (1928). On the composition of elementary errors: First paper: Mathematical deductions. *Scandinavian Actuarial Journal 1928*(1).

Cuparić, M. and B. Milošević (2024). IPCW approach for testing independence. *Journal of Nonparametric Statistics 36*(1), 118–145.

Cuparić, M., B. Milošević, and M. Obradović (2022). Asymptotic distribution of certain degenerate V- and U-statistics with estimated parameters. *Mathematical Communications 27*(1), 77–100.

De Wet, T. and R. H. Randles (1987). On the Effect of Substituting Parameter Estimators in Limiting χ^2 U and V Statistics. *The Annals of Statistics 15*(1), 398–412.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological) 39*(1), 1–22.

Diggle, P. J. (1989). Testing for random dropouts in repeated measurement data. *Biometrics 45*(4), 1255–1258.

Doretti, M., S. Geneletti, and E. Stanghellini (2018). Missing data: a unified taxonomy guided by conditional independence. *International Statistical Review 86*(2), 189–204.

Ebner, B. and N. Henze (2020). Tests for multivariate normality — A critical review with emphasis on weighted L^2-statistics. *Test 29*(4), 845–892.

Ebner, B. and N. Henze (2023). On the eigenvalues associated with the limit null distribution of the Epps-Pulley test of normality. *Statistical Papers 64*(3), 739–752.

Ebner, B., N. Henze, and S. Meintanis (2024). A unified approach to goodness-of-fit testing for spherical and hyperspherical data. *Statistical Papers 65*, 3447—3475.

Ebner, B., N. Henze, and D. Strieder (2022). Testing normality in any dimension by Fourier methods in a multivariate Stein equation. *Canadian Journal of Statistics 50*(3), 992–1033.

Ejsmont, W., B. Milošević, and M. Obradović (2023). A test for normality and independence based on characteristic function. *Statistical Papers 64*(6), 1861–1889.

Enders, C. K. (2022). *Applied Missing Data Analysis, Second Edition*. New York: Guilford Publications.

Enders, C. K. (2023). Missing data: An update on the state of the art. *Psychological Methods 30*(2), 322.

Fairclough, D. L. (2002). Design and analysis of quality of life studies in clinical trials. Technical report, CRC press.

Fernández, V. A., M. J. Gamero, and J. M. Garcia (2008). A test for the two-sample problem based on empirical characteristic functions. *Computational Statistics & Data Analysis 52*(7), 3730–3748.

Fielding, S., P. M. Fayers, and C. R. Ramsay (2009). Investigating the missing data mechanism in quality of life outcomes: a comparison of approaches. *Health and Quality of Life Outcomes 7*, 1–10.

Fischer, M. and C. Köck (2012). Constructing and generalizing given multivariate copulas: A unifying approach. *Statistics 46*(1), 1–12.

Fisher, R. A. (1925). *Statistical Methods for Research Workers*. London: Oliver and Boyd.

Franzin, A., F. Sambo, and B. Di Camillo (2017). bnstruct: an R package for Bayesian Network structure learning in the presence of missing data. *Bioinformatics 33*(8), 1250–1252.

Friedman, J. H. and L. C. Rafsky (1979). Multivariate generalizations of the Wald-Wolfowitz and Smirnov two-sample tests. *The Annals of Statistics 7*(4), 697–717.

Fuchs, C. (1982). Maximum likelihood estimation and model selection in contingency tables with missing data. *Journal of the American Statistical Association 77*(378), 270–278.

Galati, J. C. (2018a). A fresh look at ignorability for likelihood inference. *arXiv preprint arXiv:1811.05560*.

Galati, J. C. (2018b). What is meant by'$P(R|Y_{obs})$'? *arXiv preprint arXiv:1811.11011*.

Galati, J. C. (2018c). When is Y_{obs} missing and Y_{mis}observed? *arXiv preprint arXiv:1811.04161*.

Galati, J. C. (2019). Three issues impeding communication of statistical methodology for incomplete data. *arXiv preprint arXiv:1903.08880*.

Gelman, A., J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin (2014). *Bayesian Data Analysis, Third edition*. New York: Chapman and Hall/CRC.

Giacomini, R., D. N. Politis, and H. White (2013). A warp-speed method for conducting Monte Carlo experiments involving bootstrap estimators. *Econometric Theory 29*(3), 567–589.

González-Estrada, E., J. A. Villaseñor, and R. Acosta-Pech (2022). Shapiro-Wilk test for multivariate skew-normality. *Computational Statistics 37*(4), 1985–2001.

Gregory, G. G. (1977). Large Sample Theory for U-Statistics and Tests of Fit. *The Annals of Statistics 5*(1), 110 – 123.

Gretton, A., K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola (2012). A kernel two-sample test. *The Journal of Machine Learning Research 13*(1), 723–773.

Heinze, G., A.-L. Boulesteix, M. Kammer, T. P. Morris, I. R. White, and S. P. of the STRATOS Initiative (2024). Phases of methodological research in biostatistics—building the evidence base for new methods. *Biometrical Journal 66*(1), 2200222.

Henze, N. (2024). *Asymptotic Stochastics*. Springer Berlin, Heidelberg.

Henze, N. and T. Wagner (1997). A New Approach to the BHEP Tests for Multivariate Normality. *Journal of Multivariate Analysis 62*(1), 1–23.

Hoeffding, W. (1948). A Class of Statistics with Asymptotically Normal Distribution. *The Annals of Mathematical Statistics 19*(3), 293 – 325.

Hotelling, H. (1931). The generalization of Student's ratio. *The Annals of Mathematical Statistics 2*(3), 360–378.

Jamshidian, M. and S. Jalal (2010). Tests of homoscedasticity, normality, and missing completely at random for incomplete multivariate data. *Psychometrika 75*(4), 649–674.

Jamshidian, M. and M. Mata (2008). Postmodeling sensitivity analysis to detect the effect of missing data mechanisms. *Multivariate Behavioral Research 43*(3), 432–452.

Jamshidian, M. and J. R. Schott (2007). Testing equality of covariance matrices when data are incomplete. *Computational Statistics & Data Analysis 51*(9), 4227–4239.

Jamshidian, M. and K.-H. Yuan (2013). Data-driven sensitivity analysis to detect missing data mechanism with applications to structural equation modelling. *Journal of Statistical Computation and Simulation 83*(7), 1344–1362.

Jamshidian, M. and K.-H. Yuan (2014). Examining missing data mechanisms via homogeneity of parameters, homogeneity of distributions, and multivariate normality. *Wiley Interdisciplinary Reviews: Computational Statistics 6*(1), 56–73.

Jiménez-Gamero, M.-D. and M. Alba-Fernández (2021). A test for the geometric distribution based on linear regression of order statistics. *Mathematics and Computers in Simulation 186*, 103–123.

Jiménez-Gamero, M. D., J. Munoz-Garcıa, and R. Pino-Mejıas (2003). Bootstrapping parameter estimated degenerate U and V statistics. *Statistics & Probability Letters 61*(1), 61–70.

Karling, M. J., M. G. Genton, and S. G. Meintanis (2023). Goodness-of-fit tests for multivariate skewed distributions based on the characteristic function. *Statistics and Computing 33*(5), 99.

Kendall, M. G. (1975). *Rank Correlation Methods, Fourth Edition*. London: Charles Griffin, London.

Kim, K. H. and P. M. Bentler (2002). Tests of homogeneity of means and covariance matrices for multivariate incomplete data. *Psychometrika 67*, 609–623.

Klebanov, L. B., V. Beneš, and I. Saxl (2005). *N-distances and their applications*. Charles University in Prague, the Karolinum Press Prague, Czech Republic.

Kochar, S. C. and R. Gupta (1990). Distribution-free tests based on sub-sample extrema for testing against positive dependence. *Australian Journal of Statistics 32*(1), 45–51.

Kolmogorov, A. N. (1933). Sulla determinazione empirica di una legge didistribuzione. *Giorn. Dell'inst. Ital. Degli. Att. 4*, 89–91.

Koroljuk, V. S. and Y. V. Borovskich (2010). *Theory of U-Statistics*. Springer Dordrecht.

Kurita, E. and T. Seo (2022). Multivariate normality test based on kurtosis with two-step monotone missing data. *Journal of Multivariate Analysis 188*, 104824.

Lehmann, E. L. (1999). *Elements of large-sample theory*. New York: Springer.

Li, J. and Y. Yu (2015). A nonparametric test of missing completely at random for incomplete multivariate data. *Psychometrika 80*, 707–726.

Lin, J. C.-H. (2013). *A probability based framework for testing the missing data mechanism (PhD Thesis)*. Los Angeles: University of California.

Listing, J. and R. Schlittgen (1998). Tests if dropouts are missed at random. *Biometrical Journal: Journal of Mathematical Methods in Biosciences 40*(8), 929–935.

Little, R. J. and D. B. Rubin (1987). *Statistical Analysis with Missing Data, First Edition*. New York: John Wiley & Sons.

Little, R. J. and D. B. Rubin (2019). *Statistical Analysis With Missing Data, Third Edition*. New York: John Wiley & Sons.

Little, R. J. A. (1988). A Test of Missing Completely at Random for Multivariate Data with Missing Values. *Journal of the American Statistical Association 83*(404), 1198–1202.

Lopez-Paz, D. and M. Oquab (2016). Revisiting classifier two-sample tests. *arXiv preprint arXiv:1610.06545*.

Lukić, Ž. and B. Milošević (2024). A novel two-sample test within the space of symmetric positive definite matrix distributions and its application in finance. *Annals of the Institute of Statistical Mathematics 76*(5), 797–820.

Mahmoudi, M. (2024). Sales and satisfaction dataset, Kaggle, Accessed on: 10 November 2025.

Mann, H. B. and D. R. Whitney (1947). On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics 18*(1), 50–60.

McKnight, P. E., K. M. McKnight, S. Sidani, and A. J. Figueredo (2007). *Missing data: A Gentle Introduction*. New York: Guilford Press.

Mealli, F. and D. B. Rubin (2015). Clarifying missing at random and related definitions, and implications when coupled with exchangeability. *Biometrika 102*(4), 995–1000.

Mealli, F. and D. B. Rubin (2016). Clarifying missing at random and related definitions, and implications when coupled with exchangeability (Amendments and Corrections). *Biometrika 103*(2), 491–491.

Mecklin, C. J. and D. J. Mundfrom (2005). A Monte Carlo comparison of the Type I and Type II error rates of tests of multivariate normality. *Journal of Statistical Computation and Simulation 75*(2), 93–107.

Meintanis, S., B. Milošević, M. Obradović, and M. Veljović (2024). Goodness-of-fit tests for the multivariate Student-t distribution based on iid data, and for GARCH observations. *Journal of Time Series Analysis 45*(2), 298–319.

Meyer, C. D. (2023). *Matrix analysis and applied linear algebra*. Philadelphia: SIAM.

Mirzaei, A., S. R. Carter, A. E. Patanwala, and C. R. Schneider (2022). Missing data in surveys: Key concepts, approaches, and applications. *Research in Social and Administrative Pharmacy 18*(2), 2308–2316.

Neuhaus, G. (1977). Functional limit theorems for U-statistics in the degenerate case. *Journal of Multivariate Analysis 7*(3), 424–439.

Ofner, M., S. Hörmann, D. Kraus, and D. Liebl (2025). Testing the Missing Completely at Random Assumption for Functional Data. *arXiv preprint arXiv:2505.08721*.

Park, T. and C. S. Davis (1993). A test of the missing data mechanism for repeated categorical data. *Biometrics 49*(2), 631–638.

Park, T., S. Lee, and R. F. Woolson (1993). A test of the missing data mechanism for repeated measures data. *Communications in Statistics-Theory and Methods 22*(10), 2813–2829.

Park, T. and S.-Y. Lee (1997). A test of missing completely at random for longitudinal data with missing observations. *Statistics in Medicine 16*(16), 1859–1871.

Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science 50*(302), 157–175.

Pettitt, A. N. (1976). A two-sample Anderson-Darling rank statistic. *Biometrika 63*(1), 161–168.

Potthoff, R. F., G. E. Tudor, K. S. Pieper, and V. Hasselblad (2006). Can one assess whether missing data are missing at random in medical studies? *Statistical Methods in Medical Research 15*(3), 213–234.

Qu, A. and P. X.-K. Song (2002). Testing ignorable missingness in estimating equation approaches for longitudinal data. *Biometrika 89*(4), 841–850.

Randles, R. H. (1982). On the Asymptotic Normality of Statistics with Estimated Parameters. *The Annals of Statistics 10*(2), 462 – 474.

Ridout, M. S. and P. J. Diggle (1991). Testing for random dropouts in repeated measurement data. *Biometrics 47*(4), 1617–1621.

Rockel, T. (2023). *missMethods: Methods for Missing Data*. R package version 0.4.0.

Rouzinov, S. and A. Berchtold (2022). Regression-based approach to test missing data mechanisms. *Data 7*(2), 1–16.

Rubin, D. B. (1976). Inference and Missing Data. *Biometrika 63*(3), 581–592.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons Inc.

Santos, M. S., R. C. Pereira, A. F. Costa, J. P. Soares, J. Santos, and P. H. Abreu (2019). Generating synthetic missing data: A review by missing mechanism. *IEEE Access 7*, 11651–11667.

Seaman, S., J. Galati, D. Jackson, and J. Carlin (2013). What Is Meant by "Missing at Random"? *Statistical Science 28*(2), 257 – 268.

Sejdinovic, D., B. Sriperumbudur, A. Gretton, and K. Fukumizu (2013). Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *Annals of Statistics*, 2263–2291.

Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics*. New York: John Wiley & Sons.

Simon, B. (2005). *Trace ideals and their applications*. Providence: American Mathematical Soc.

Sklar, A. (1959). Fonctions de répartition à n dimensions et leurs marges. *Annales de l'ISUP 8*(3), 229–231.

Smirnov, N. V. (1939). On the estimation of the discrepancy between empirical curves of distribution for two independent samples. *Bull. Math. Univ. Moscou 2*(2), 3–14.

Spohn, M.-L., J. Näf, L. Michel, and N. Meinshausen (2021). PKLM: A flexible MCAR test using Classification. *arXiv preprint arXiv:2109.10150*.

Stekhoven, D. J. (2013). Package 'missforest'. *R package version 1*, 21.

Strang, G. (2016). *Introduction to Linear Algebra (5th Edition)*. Wellesley-Cambridge Press.

Student (1908). The probable error of a mean. *Biometrika 6*(1), 1–25.

Székely, G. J. and M. L. Rizzo (2004). Testing for equal distributions in high dimension. *InterStat 5*(16.10), 1249–1272.

Székely, G. J. and M. L. Rizzo (2013). Energy statistics: A class of statistics based on distances. *Journal of Statistical Planning and Inference 143*(8), 1249–1272.

Tan, M., H.-B. Fang, G.-L. Tian, and G. Wei (2005). Testing multivariate normality in incomplete data of small sample size. *Journal of Multivariate Analysis 93*(1), 164–179.

Tsatsi, A., A. Batsidis, and P. Economou (2024). Multivariate normality tests with two-step monotone missing data: a critical review with emphasis on the different methods of handling missing values. *Journal of Statistical Computation and Simulation 94*(16), 3653–3677.

Van Buuren, S. (2018). *Flexible Imputation of Missing Data, Second Edition*. New York: Chapman and Hall/CRC.

von Mises, R. (1928). *Wahrscheinlichkeit Statistik und Wahrheit*. Springer Berlin, Heidelberg.

von Mises, R. (1947). On the asymptotic distribution of differentiable statistical functions. *The Annals of Mathematical Statistics 18*(3), 309–348.

Wald, A. and J. Wolfowitz (1940). On a test whether two samples are from the same population. *The Annals of Mathematical Statistics 11*(2), 147–162.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin 1*(6), 80–83.

Yamada, T., M. M. Romer, and D. S. P. Richards (2015). Kurtosis tests for multivariate normality with monotone incomplete data. *Test 24*(3), 532–557.

Yuan, K.-H., M. Jamshidian, and Y. Kano (2018). Missing data mechanisms and homogeneity of means and variances–covariances. *Psychometrika 83*, 425–442.

Zeng, Y., N. M. Adams, and D. A. Bodenham (2024). Mmd two-sample testing in the presence of arbitrarily missing data. *arXiv preprint arXiv:2405.15531*.

Zhang, S., P. Han, and C. Wu (2019). A unified empirical likelihood approach for testing MCAR and subsequent estimation. *Scandinavian Journal of Statistics 46*(1), 272–288.

Biography

Danijel G. Aleksić was born on September 25, 1998, in Zvornik, Republic of Srpska, Bosnia and Herzegovina, to Gojko Aleksić (born 1967), a truck driver, and Milofinka Aleksić (née Kapetanović, 1971–2021), a housewife. He is an only child.

He completed the first grade of elementary school at the Elementary School "Sveti Sava" in Zelinje, Zvornik. He attended the second, third, and fourth grades at "Vuk Karadžić" Elementary School in Roćević, Zvornik, and completed the remainder of his elementary education at the branch of the same school in Branjevo, Zvornik. He graduated elementary school with a GPA of 4.93/5.00. He participated in, and sometimes won, various municipal and regional competitions in mathematics, English, physics, and chemistry.

He completed his secondary education at "Vuk Karadžić" Grammar School in Loznica, Serbia, specializing in the natural sciences and mathematics, graduating with the "Vuk Karadžić" diploma. He also participated in, and occasionally won, municipal and regional competitions in mathematics and Russian.

In 2017, he enrolled at the Faculty of Mathematics, University of Belgrade, in Mathematics program, with a focus on Statistics, Actuarial and Financial Mathematics. He graduated on September 21, 2021, with a GPA of 9.47/10.00. In the same year, he enrolled in the master's program in the same field, which he completed one year later, on September 22, 2022, with a GPA of 8.75/10.00. He defended his master's thesis entitled *"The Problem of Missing Data: Impact on Statistical Inference"* (in Serbian), under the supervision of professor Bojana Milošević, PhD. That same year, he enrolled in the doctoral academic studies in the same field, where he passed all required exams with a GPA of 9.88/10.00.

During the 2021/2022 academic year, he worked as a teaching associate in the Department of Probability and Statistics at the Faculty of Mathematics, University of Belgrade. In the 2022/2023 academic year, he held the same position in the Department of Mathematics at the Faculty of Organizational Sciences, where he has been employed as a teaching assistant since November 2023. In the 2024/2025 academic year, he also worked part-time as a teacher of Analysis with Algebra at the Mathematical Grammar School in Belgrade.

His primary research interests include the problem of missing data, particularly the development of new tests for the MCAR assumption, as well as the adaptation of existing statistical procedures in the presence of missing data.

He is proficient in English and speaks Russian at an intermediate level.

Прилог 1.

Изјава о ауторству

Потписани-а Данијел Алексић

број уписа 2006/2022

Изјављујем

да је докторска дисертација под насловом

U- and V-statistics for incomplete data and their application to model specification testing (U- и V-статистике за некомплетне податке и њихова примена у тестирању сагласности са моделом)

- резултат сопственог истраживачког рада,
- да предложена дисертација у целини ни у деловима није била предложена за добијање било које дипломе према студијским програмима других високошколских установа,
- да су резултати коректно наведени и
- да нисам кршио/ла ауторска права и користио интелектуалну својину других лица.

Потпис докторанда

У Београду, 21.01.2026.

Прилог 2.

Изјава о истоветности штампане и електронске верзије докторског рада

Име и презиме аутора Данијел Алексић

Број уписа 2006/2022

Студијски програм Математика

Наслов рада U- and V-statistics for incomplete data and their application to model specification testing (U- и V-статистике за некомплетне податке и њихова примена у тестирању сагласности са моделом)

Ментор проф. др Бојана Милошевић

Потписани Данијел Алексић

изјављујем да је штампана верзија мог докторског рада истоветна електронској верзији коју сам предао/ла за објављивање на порталу **Дигиталног репозиторијума Универзитета у Београду.**

Дозвољавам да се објаве моји лични подаци везани за добијање академског звања доктора наука, као што су име и презиме, година и место рођења и датум одбране рада.

Ови лични подаци могу се објавити на мрежним страницама дигиталне библиотеке, у електронском каталогу и у публикацијама Универзитета у Београду.

Потпис докторанда

У Београду, 21.01.2026.

Прилог 3.

Изјава о коришћењу

Овлашћујем Универзитетску библиотеку „Светозар Марковић“ да у Дигитални репозиторијум Универзитета у Београду унесе моју докторску дисертацију под насловом:

U- and V-statistics for incomplete data and their application to model specification testing (U- и V-статистике за некомплетне податке и њихова примена у тестирању сагласности са моделом)

која је моје ауторско дело.

Дисертацију са свим прилозима предао/ла сам у електронском формату погодном за трајно архивирање.

Моју докторску дисертацију похрањену у Дигитални репозиторијум Универзитета у Београду могу да користе сви који поштују одредбе садржане у одабраном типу лиценце Креативне заједнице (Creative Commons) за коју сам се одлучио/ла.

1. Ауторство

2. Ауторство - некомерцијално

3. Ауторство – некомерцијално – без прераде

4. Ауторство – некомерцијално – делити под истим условима

5. Ауторство – без прераде

6. Ауторство – делити под истим условима

(Молимо да заокружите само једну од шест понуђених лиценци, кратак опис лиценци дат је на полеђини листа).

Потпис докторанда

У Београду, 21.01.2026.

1. Ауторство - Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, и прераде, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце, чак и у комерцијалне сврхе. Ово је најслободнија од свих лиценци.

2. Ауторство – некомерцијално. Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, и прераде, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце. Ова лиценца не дозвољава комерцијалну употребу дела.

3. Ауторство - некомерцијално – без прераде. Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, без промена, преобликовања или употребе дела у свом делу, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце. Ова лиценца не дозвољава комерцијалну употребу дела. У односу на све остале лиценце, овом лиценцом се ограничава највећи обим права коришћења дела.

4. Ауторство - некомерцијално – делити под истим условима. Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, и прераде, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце и ако се прерада дистрибуира под истом или сличном лиценцом. Ова лиценца не дозвољава комерцијалну употребу дела и прерада.

5. Ауторство – без прераде. Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, без промена, преобликовања или употребе дела у свом делу, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце. Ова лиценца дозвољава комерцијалну употребу дела.

6. Ауторство - делити под истим условима. Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, и прераде, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце и ако се прерада дистрибуира под истом или сличном лиценцом. Ова лиценца дозвољава комерцијалну употребу дела и прерада. Слична је софтверским лиценцама, односно лиценцама отвореног кода.