

This is a “preproof” accepted article for *Psychometrika*.

This version may be subject to change during the production process.

DOI: 10.1017/psy.2024.14

PKLM: A FLEXIBLE MCAR TEST USING CLASSIFICATION

META-LINA SPOHN¹ *

JEFFREY NÄF² *

LORIS MICHEL³

NICOLAI MEINSHAUSEN¹

¹ETH ZURICH; ²INRIA PREMEDICAL TEAM; ³QUANTCO

September 15, 2024

Acknowledgement We are grateful to Jun Li for providing us with parts of their code.

*Authors with equal contribution.

Correspondence should be sent to

E-Mail:

jeffrey.naf@inria.fr

This is an Open Access article, distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives licence (<http://creativecommons.org/licenses/by-nc-nd/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is unaltered and is properly cited. The written permission of Cambridge University Press must be obtained for commercial re-use or in order to create a derivative work.

Abstract

We develop a fully non-parametric, easy-to-use, and powerful test for the missing completely at random (MCAR) assumption on the missingness mechanism of a dataset. The test compares distributions of different missing patterns on random projections in the variable space of the data. The distributional differences are measured with the Kullback-Leibler Divergence, using probability Random Forests (Malley et al., 2011). We thus refer to it as “Projected Kullback-Leibler MCAR” (PKLM) test. The use of random projections makes it applicable even if very few or no fully observed observations are available or if the number of dimensions is large. An efficient permutation approach guarantees the level for any finite sample size, resolving a major shortcoming of most other available tests. Moreover, the test can be used on both discrete and continuous data. We show empirically on a range of simulated data distributions and real datasets that our test has consistently high power and is able to avoid inflated type-I errors. Finally, we provide an R-package PKLMtest with an implementation of our test.

Key words: Random Projections, Tree Ensembles, Random Forest, KL-Divergence, Permutation

1. Introduction

Dealing with missing values is an integral part of modern statistical analysis. In particular, the assumed mechanism leading to the missing values is of great importance. Based on the work of Rubin (1976), there are three groups of missingness mechanisms usually considered: The values may be missing completely at random (MCAR), meaning the probability of a value being missing does not depend on the observed or unobserved data. In contrast, the probability of being missing could depend on observed values (missing at random, MAR) or on unobserved values (missing not at random, MNAR).

As stated in Yuan et al. (2018), “a formal confirmation of the MCAR missing data mechanism is of great interest, simply because essentially all methods can still yield consistent estimates under MCAR even if the underlying population distribution is unknown”. While there is, at least for imputation, a number of approaches that can deal with a MAR missing data mechanism such as Multivariate Imputation by Chained Equations (mice) (Buuren and Groothuis-Oudshoorn, 2011; Deng et al., 2016), many commonly used methods explicitly rely on the validity of the MCAR assumption. Examples are the easy-to-use listwise-deletion and mean-imputation methods (Little and Rubin, 1986). Consequently, the original paper on MCAR testing (Little, 1988) has been cited close to 10'000 times according to google scholar. Recent papers (involving psychometric analysis) that test the MCAR assumption in order to justify listwise-deletion include Brown et al. (2020), Charles et al. (2021), Hawes et al. (2021), Sun and Chen (2022), de Vos et al. (2022), Rajeb et al. (2023), Zarate et al. (2023), and Langer et al. (2024). As such, it is important to reliably test the MCAR assumption.

The testing framework is of an ANOVA-type: when observing a dataset with missing values, there are n observations and G missingness patterns, $g = 1, \dots, G$. The observations belonging to the missingness pattern g can be seen as a group, such that we observe G groups of observations. The MCAR hypothesis now implies that the distribution of the observed data in all groups is the same, while under the alternative at least two differ. This is technically testing the observed at random (OAR) assumption defined in Rhoads (2012), see also the end of Section 3 for a discussion. This distinction can be avoided by assuming the missingness mechanism is MAR,

which is what is usually implicitly done (Li and Yu, 2015).

The idea of testing the MCAR assumption traces back to Little (1988). While some more refined versions of this testing idea were developed since then (Chen and Little, 1999; Kim and Bentler, 2002; Jamshidian and Jalal, 2010), there has not been a lot of progress on distribution-free MCAR tests, able to detect general distributional differences between the missingness patterns. Li and Yu (2015) recently made a step in that direction. Their test is completely nonparametric and shown to be consistent. Empirically it is shown to keep the level and to have a high power over a wide range of distributions. An application area where their proposed test struggles is for higher-dimensional data with little or no complete observations. Their testing paradigm is based on “a reasonable amount” of complete cases and all pairwise comparisons between the observed parts of two missingness pattern groups. This is problematic, since, as the dimension p increases, the number of distinct patterns G tends to grow quickly as well. The most extreme case occurs when $G = n$, that is, every observation forms a missingness pattern group on its own. Consequently, their test appears computationally prohibitively expensive for $p > 10$. Additionally, as the dimension increases, both the number of complete cases and the number of observations per pattern tends to decrease, both contributing to a reduction in power for the test in Li and Yu (2015).

In this paper, we try to circumvent these problems in a data-efficient way, by employing a one v.s. all-others approach and using *random projections* in the variable space. Considering observations that are projected into a lower-dimensional space allows us to recover more complete cases. As realized by Li and Yu (2015), the problem of MCAR testing, as described above, is a problem of testing whether distributions across missingness patterns are different. The method presented here relies on some of the core ideas of Näf et al. (2023) and Cai et al. (2020), who do distributional testing using classifiers. We extend the ideas of Cai et al. (2020) to be usable for multiclass classification and use the projection idea of Näf et al. (2023) to build a test that is usable and powerful even for high dimensions. Moreover, using a permutation approach, we are able to provably keep the nominal level α for all n . As outlined later, this is in contrast to other tests, for which the level might be kept only asymptotically, or is even unclear. The approach of random projections together with a permutation test also allows to extract more information than

just a global hypothesis test. We make use of this to calculate individual p -values for each variable. Such a partial test for a variable addresses the null hypothesis that, once that variable is removed, the data is MCAR. Together with the test of overall MCAR, this might point towards the potential source of deviation from the null, that is, the variables causing an MCAR violation.

The paper is structured in the following way. Section 2 introduces notation. Section 3 details the testing framework including the null and alternative hypotheses we consider. Section 4 then showcases how to perform this test in practice and details the algorithm. Section 5 shows some numerical comparisons for type-I error control and power. Section 6 explains the extension of partial p -values, while Section 7 concludes. Appendix A contains the proofs of all results, while Appendix B adds some additional details and shows computation times of the different tests.

1.1. Contributions

Our contributions can be summarized as follows: We develop the PKLM-test, an easy-to-use and powerful non-parametric test for MCAR, that is applicable even in high dimensions. We thereby extend the testing approach of Cai et al. (2020) to multiclass testing, which in connection with random projections in the variable space and the Random Forest classifier leads to a powerful test for both discrete and continuous types of data. To the best of your knowledge, no other test is as widely applicable and powerful. Moreover, we are able to formally prove the validity of our p -values for any sample size and number of groups G . As we demonstrate in our simulations, this is remarkable for the MCAR testing literature. It appears no other MCAR test has such a guarantee and many have inflated type-I errors, even in realistic cases, see e.g. the discussion in Jamshidian and Jalal (2010).

As an extension, we can compute partial p -values corresponding to each variable, addressing the question of the source of violation of MCAR among the variables. We demonstrate the validity and power of our test on a wide range of simulated and real datasets in conjunction with different MAR mechanisms. Finally, we make our test available through the R-package `PKLMtest`, available on <https://github.com/missValTeam/PKLMtest> and on CRAN.

1.2. Related Work

Previous advances for tests of MCAR were mostly addressed by [Little \(1988\)](#) (referred to as “Little-test”) and extensions ([Chen and Little, 1999](#); [Kim and Bentler, 2002](#)) under the assumption of joint Gaussianity. To the best of our knowledge, the only distribution-free tests are developed in [Jamshidian and Jalal \(2010\)](#), [Li and Yu \(2015\)](#) and [Zhang et al. \(2019\)](#). The first paper develops a test (referred to as “JJ-test”), which is distribution-free but is only able to spot differences in the covariance matrices between the different patterns. As such, the simulation study in [Li and Yu \(2015\)](#) shows that their test (referred to as “Q-test”), which can detect any potential difference, has much more power than the JJ-test. Moreover, the JJ-test requires prior imputation of missing values, which appears undesirable. [Zhang et al. \(2019\)](#) develop a test that can be used to subsequently also consistently estimate certain estimators under MCAR. Their test requires a set of fully observed “auxiliary” variables that can be used to first test and then estimate properties of some variable of interest. As such their approach and goals are quite different from ours.

Consequently, the test closest to ours is the fully non-parametric method in [Li and Yu \(2015\)](#). However, it is computationally costly or even infeasible to use their test with dimensions typically found in modern datasets ($p \gg 10$), as all pairwise comparisons between missingness patterns are calculated. While this could in principle be avoided by only checking a subset of pairs, we empirically show that, even if all pairwise comparisons are performed, our test has comparable or even higher power than theirs in their own simulation setting. This gap only increases with the number of dimensions or with a decrease in the fraction of fully observed cases.

We also address a major issue in the MCAR testing literature: none of the proposed methods has a finite sample guarantee of producing valid p -values and for some it can even be empirically checked that the produced p -value is not valid in certain settings. If Z is a p -value generated from a statistical test, then it is valid if $\mathbb{P}(Z \leq \alpha) \leq \alpha$ under H_0 for all $\alpha \in [0, 1]$, see e.g., [Lehmann and Romano \(2005\)](#). [Figure 2](#) in [Section 5](#) shows some example of previous tests violating this validity of p -values. This issue might be surprising since the requirement of a valid p -value might be the most basic demand a statistical test needs to meet. For the Little-test, this is generally

true under normality or asymptotically, that is if the number of observations is going to infinity, under some moment conditions and conditions on the group size. Despite this, Section 5 shows that type error rates can strongly exceed the desired level even in samples of 500 observations. The same holds for the JJ-test of [Jamshidian and Jalal \(2010\)](#) for which we sometimes observed a strong inflation of the level. As with the JJ-test, [Li and Yu \(2015\)](#) also do not provide a formal guarantee that the level is kept. Though in our own simulation study, which is similar to theirs, we did not find any notable violation of the level for their test.

To conduct our test, we adapt and partially extend the approaches of [Cai et al. \(2020\)](#) and [Näf et al. \(2023\)](#). The former develops a two-sample test using classification, an approach that has gained a lot of attention in recent years (see e.g., [Kim et al. \(2021\)](#) or [Hediger et al. \(2022\)](#) for a literature overview). We extend this approach to multiclass testing, to obtain a test statistic akin to [Cai et al. \(2020\)](#), but using the out of bag (OOB) probability estimate of the Random Forest (RF) instead of the in-sample probability. This was already hinted in [Hediger et al. \(2022\)](#) to increase the power of the two-sample testing approach designed by [Cai et al. \(2020\)](#). [Näf et al. \(2023\)](#), on the other hand, use random projections to increase the sample efficiency in the presence of missing values. This simple idea makes our test applicable and powerful, even in high dimensions, and even if the number of patterns G is the same as the number of observations. It can also provide additional information together with the rejection decision, as we demonstrate in Section 6. Finally, through an efficient permutation testing approach, we are able to formally guarantee that our test produces valid p -values for *any* n and any number of groups G . It appears that the PKLM-test is the first MCAR test with such a guarantee. Table 1 summarizes some of the properties of different tests. In particular, “mixed data types” refers to a possible combination of continuous data (such as income) and discrete data (such as gender), while “power beyond differences in first and second moments” means the test is able to detect differences between distributions, even if their means or variances are identical. Though this is difficult to show formally, it appears quite clear that the nonparametric nature of our approach allows for the detection of differences in distributions between patterns, even if the missingness groups all share the same mean or covariance matrix. As outlined in [Yuan et al. \(2018\)](#) this is crucial for the detection of general MCAR deviations and is not the case, for instance, for the widely used

Little-test. Appendix C studies a simulated MAR example taken from Yuan et al. (2018), whereby observed means and variances are approximately the same across different groups. Tests such as the Little-test have no power in this example, yet with our approach, we reach a power of 1.

	PKLM	Q	Little	JJ
Computational Complexity	$\mathcal{O}(pn \log(n))$	$\mathcal{O}(n^2p)$	$\mathcal{O}(np^2)$	$\mathcal{O}(n(p^2 + \log(n)))$
Can be used without complete observations	Yes	No	No	Yes
Mixed data types possible	Yes	No	No	No
Does not require initial imputation	Yes	Yes	Yes	No
Power beyond differences in first and second moments	Yes	Yes	No	No

Table 1: Illustration of some of the properties of various tests. For details on the calculation of the computational complexities we refer to Appendix B.

2. Notation

We assume an underlying probability space $(\Omega, \mathcal{F}, \mathbb{P})$ on which all random elements are defined. Along the lines of Muzellec et al. (2020) we introduce the following notation: let $\mathbf{X}^* \in \mathbb{R}^{n \times p}$ be a matrix of n complete samples from a distribution P^* on \mathbb{R}^p . We denote by \mathbf{X} the corresponding incomplete dataset that is actually observed. Alongside \mathbf{X} we observe the missingness matrix $\mathbf{M} \in \{0, 1\}^{n \times p}$, of which an entry $m_{ij} \in \{0, 1\}$ is 1, if entry x_{ij}^* is missing, and 0, if it is observed. Each unique combination in $\{0, 1\}^p$ in \mathbf{M} is referred to as a missingness pattern and we assume that there are $G \leq n$ unique patterns in \mathbf{M} . As an example, for $p = 2$, we might have the pattern $(1, 0)$ (first value missing, second observed), $(0, 1)$ (first value observed, second missing) or $(0, 0)$ (both values are observed). We do not consider the completely missing pattern, in this case $(1, 1)$.

We assume that each row x_i (x_i^*) of \mathbf{X} (\mathbf{X}^*) is a realization of an i.i.d. copy of the random vector X (X^*) with distribution P (P^*). Similarly, M is the random vector in $\{0, 1\}^p$ encoding the missingness pattern of X . Furthermore we assume that P (P^*) has a density f (f^*) with

respect to some dominating measure. For a random vector X or an observation x in \mathbb{R}^p and subset $A \subseteq \{1, \dots, p\}$, we denote as X_A (x_A) the projection onto that subset of indices. For instance if $p = 3$ and $A = \{1, 2\}$, then $X_A = (X_1, X_2)$ ($x_A = (x_1, x_2)$). For any set $C \subseteq \{1, \dots, p\}$, we denote by $\mathbf{X}_{\bullet C}$ the matrix of n observations projected onto dimensions in C , so that $\mathbf{X}_{\bullet C}$ is of dimension $n \times |C|$. Similarly, for $R \subseteq \{1, \dots, n\}$, $\mathbf{X}_{R\bullet}$ denotes the matrix of observations in set R , over all dimensions, so that the dimension of $\mathbf{X}_{R\bullet}$ is given by $|R| \times p$. We denote by F_g (respectively f_g) the complete distribution (density) of the data in the g^{th} missingness pattern group. A quick overview of the notation including the use of indices for the number of missingness patterns, dimensions, observations, projections and permutations is given in Table 2.

notation	partial	full
distribution	P	P^*
dataset	\mathbf{X}	\mathbf{X}^*
observation in \mathbb{R}^p	x_i	x_i^*
random vector	X	X^*
density	f	f^*
number of missingness patterns	G	
number of dimensions	p	
number of observations	n	
number of projections	N	
number of permutations	L	

Table 2: Notation: Summary of the notation used throughout the paper, with (“partial”) and without (“full”) considering the missing values.

3. Testing Framework

In this section, we formulate the specific null and alternative hypotheses for testing MCAR considered by the PKLM-test. Recalling the notation of Section 2, a missingness pattern is

defined by a vector of length p , consisting of ones and zeros, indicating which of the p variables are missing in the given pattern. We divide the n observations into $g \in \{1, \dots, G\}$ unique groups, such that the observations of each group share the same missingness pattern. Each group $g \in \{1, \dots, G\}$ contains n_g observations such that $n_1 + \dots + n_G = n$. Let F_g denote the joint distribution of the p variables in the missingness pattern group g , such that the n_g observations of the group g are i.i.d. draws from F_g . As stated in Li and Yu (2015), testing MCAR can be formulated by the hypothesis testing problem

$$\begin{aligned} H_0 : F_1^* = F_2^* = \dots = F_G^* \\ \text{v.s.} \\ H_A : \exists i \neq j \in \{1, \dots, G\} \text{ s.t. } F_i^* \neq F_j^*. \end{aligned} \tag{1}$$

We want to emphasize the use of F^* in the testing problem (1), indicating that these hypotheses involve distributions we cannot access. Thus, (1) needs to be weakened. Borrowing the notation of Li and Yu (2015), for missingness pattern group g we denote with \mathbf{o}_g and \mathbf{m}_g the subsets of $\{1, \dots, p\}$ indicating which variables are observed and which are missing, respectively. We denote the induced distributions by F_{g, \mathbf{o}_g} and F_{g, \mathbf{m}_g} . For two groups i and j , we denote by $\mathbf{o}_{ij} := \mathbf{o}_i \cap \mathbf{o}_j$ the shared observed variables of both groups. As mentioned in Li and Yu (2015), it is not possible to test (1) reliably, since the distribution F_{i, \mathbf{m}_i} of the unobserved variables is inaccessible. Thus, Li and Yu (2015) consider the following hypothesis testing problem

$$\begin{aligned} H_0 : F_{i, \mathbf{o}_{ij}} = F_{j, \mathbf{o}_{ij}} \quad \forall i \neq j \in \{1, \dots, G\} \\ \text{v.s.} \\ H_A : \exists i \neq j \in \{1, \dots, G\} \text{ with } \mathbf{o}_{ij} \neq \emptyset \text{ s.t. } F_{i, \mathbf{o}_{ij}} \neq F_{j, \mathbf{o}_{ij}}. \end{aligned} \tag{2}$$

The null hypothesis H_0 of (2) is implied by H_0 of (1), but not vice-versa. In other words, if we can reject the null hypothesis of (2), we can also reject the null hypothesis of (1). But if the null hypothesis of (2) cannot be rejected, there could still be a distributional change for different groups in the unobserved parts, so that the null hypothesis of (1) is not true. In this case, the missingness mechanism would be MNAR. Thus, using the terminology of Rhoads (2012), (2) tests

the “observed at random” (OAR) hypothesis instead of the MCAR hypothesis. The differentiation can be circumvented by assuming that the missingness mechanism is MAR, which is the approach usually taken, see [Li and Yu \(2015\)](#).

The comparison of all pairs of missingness groups in the hypothesis testing problem (2) is problematic however, as laid out in the introduction. In the following, we circumvent this problem in a data-efficient way, considering a one v.s. all-others approach and employing *random projections* in the variable space. Considering observations that are projected into a lower-dimensional space allows us to recover more complete cases. Let \mathcal{A} be the set of all possible subsets of $\{1, \dots, p\}$ with at most $p - 1$ elements. For $A \in \mathcal{A}$ we define by \mathcal{N}_A the indices in $1, \dots, n$ of observations that are observed with respect to projection A , i.e., observations of which the projection onto A is fully observed. These observations may belong to different missingness pattern groups $g \in \{1, \dots, G\}$. As an example, $x = (\text{NA}, 1, \text{NA}, 2, 4)$ and $y = (\text{NA}, \text{NA}, \text{NA}, 1, 3)$ are not complete and not in the same group, however if we project them to the dimensions $A = \{4, 5\}$, x_A and y_A are complete in this lower-dimensional space.

Additionally, to circumvent the problem of many groups with only a few members, we assign new *grouping or class labels* to all observations in \mathcal{N}_A . To do so, we consider the set of projections $\mathcal{B}(A^c)$, which is defined as the power set of $\{1, \dots, p\} \setminus A$. The set $\mathcal{B}(A^c)$ is never empty since $|A| \leq p - 1$. For a given projection $B \in \mathcal{B}(A^c)$, we project all observations with index in \mathcal{N}_A to B and form new collapsed missingness pattern groups $G(A, B)$, where $G(A, B)$ is the set of labels corresponding to distinct missingness patterns among observations with index in \mathcal{N}_A projected to B . This is solely done to determine the grouping or class labels of observations with index in \mathcal{N}_A . If two observations with index in \mathcal{N}_A are in the same overall missingness pattern group $g \in \{1, \dots, G\}$, they also end up in the same collapsed group. The other direction is not true, that is the number of collapsed groups $|G(A, B)|$ is at most as large as the initial number of distinct groups G among the observations with index in \mathcal{N}_A . Considering again $x = (\text{NA}, 1, \text{NA}, 2, 4)$ and $y = (\text{NA}, \text{NA}, \text{NA}, 1, 3)$, if $B = \{1, 2\}$, then observations x and y are not in the same missingness pattern group. However, if $B = \{1, 3\}$, we assign the same class label to x and y . Thus, given the projection A , we obtain a set of fully observed observations $\mathbf{X}_{\mathcal{N}_A, A} = \mathbf{X}_{\mathcal{N}_A, A}^*$, and given the projection B we assign to them the $|G(A, B)|$ different class

labels. Figure 1 provides a schematic illustration of projections A and B on a more complicated example with four observations, each corresponding to a different pattern (i.e., $n = G = 4$).

According to $B = \{2\}$, the first observation in $\mathbf{X}_{\mathcal{N}_A, A}$ obtains one collapsed class label whereas the second and third observation obtain another, common label, resulting in $|G(A, B)| = 2$.

We are now equipped to formulate our one v.s. all-others approach with the hypothesis testing problem

$$\begin{aligned}
 H_0 : F_{g,A} &= \sum_{j \in G(A,B) \setminus g} \omega_j^g F_{j,A} \\
 \forall g \in G(A, B), \forall B \in \mathcal{B}(A^c), \forall A \in \mathcal{A} \\
 \text{v.s.} & \\
 H_A : F_{g,A} &\neq \sum_{j \in G(A,B) \setminus g} \omega_j^g F_{j,A}. \\
 &\text{for one } g \in G(A, B), B \in \mathcal{B}(A^c), A \in \mathcal{A},
 \end{aligned} \tag{3}$$

where $F_{g,A}$ is the joint distribution of the observations of class g with index in \mathcal{N}_A and the groups $j \in G(A, B)$ are jointly determined by A and B . Thus, we compare the distribution of the observed part with respect to A of one group g with the mixture of the observed parts of the rest of the groups. The weights ω_j^g are non-negative, sum to 1, and are proportional to the respective fraction of observations in class j .

Example 1. To give some intuition about the hypothesis testing problem (3), we relate it to the hypothesis testing problem (2) with the help of the example of Figure 1. In this example, each observation $i = 1, \dots, 4$ has a different pattern and can thus be seen as a draw from a distribution F_i^* . We first assume that the null hypothesis of (3) holds and show, as an example, that this implies $F_{1, \mathbf{o}_{13}} = F_{3, \mathbf{o}_{13}}$. Since the null hypothesis of (3) refers to all $A \in \mathcal{A}$, it also includes $A = \mathbf{o}_{13} = \{3, 4, 5\}$, which is what we consider in Figure 1. While we are only interested in $F_{1,A}$ and $F_{3,A}$, taking $B = \{1, 2\}$ the observations in $\mathbf{X}_{\mathcal{N}_A, A}$ come from the three distributions

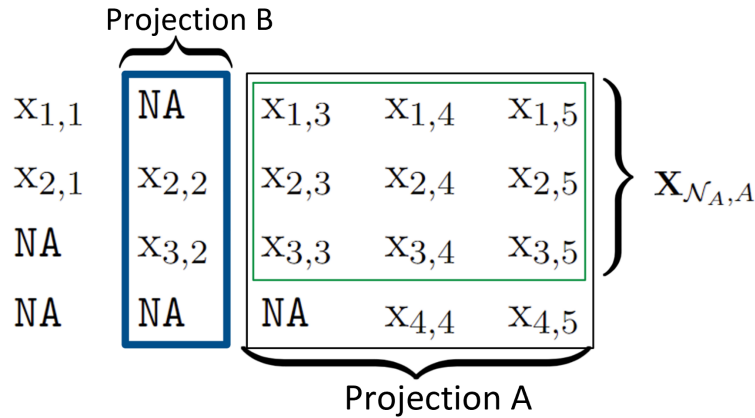


Figure 1: Illustration of the projections A and B in an example with $n = 4$ and $p = 5$. In a first step, a projection $A = \{3, 4, 5\} \subset \{1, \dots, 5\}$ is drawn. The fully observed points on A form $\mathbf{X}_{\mathcal{N}_A, A}$, as indicated in green. In a second step, a projection $B = \{2\} \subset \{1, \dots, 5\} \setminus A$ is drawn, as indicated in blue. The patterns in projection B then determine the labels assigned to the observations in $\mathbf{X}_{\mathcal{N}_A, A}$. In this case we obtain two different class labels: the first observation has one label, and the second and third observations share another common label.

$F_{1,A}, F_{2,A}, F_{3,A}$. Due to (3) it holds that

$$\begin{aligned}
 F_{1,A} &= \omega_2^1 F_{2,A} + \omega_3^1 F_{3,A}, \\
 F_{2,A} &= \omega_1^2 F_{1,A} + \omega_3^2 F_{3,A}, \\
 F_{3,A} &= \omega_1^3 F_{1,A} + \omega_2^3 F_{2,A}.
 \end{aligned}
 \tag{4}$$

Some algebra shows that equation system (4) is equivalent to $F_{1,A} = F_{2,A} = F_{3,A}$, which in particular means $F_{1,A} = F_{3,A}$, that we wanted to show. While we took $i = 1$ and $j = 3$ as an example matching Figure 1, we cycle through all $A \in \mathcal{A}$ in (3) and thus $A = \mathbf{o}_{ij}$ for all patterns i, j eventually. We now assume that the null hypothesis of (2) is true and consider again $A = \{3, 4, 5\}$ as an example. Since we only look at the fully observed observations in \mathcal{N}_A in (3), i.e., leave out the fourth point, we again deal with the three distributions $F_{1,A}, F_{2,A}, F_{3,A}$. Moreover, by construction, $A \subset \mathbf{o}_{12}$ and $A \subset \mathbf{o}_{13}$ (even $A = \mathbf{o}_{13}$ in this case). Thus, $F_{1,\mathbf{o}_{12}} = F_{2,\mathbf{o}_{12}}$ and $F_{1,\mathbf{o}_{13}} = F_{3,\mathbf{o}_{13}}$, implied by the null hypothesis of (2), means that $F_{1,A} = F_{2,A} = F_{3,A}$, which implies (4). Again this might seem constructed, but since by definition,

(3) only considers the distributions $F_{i,A}$ and $F_{j,A}$ of fully observed points on A , it will always hold that $A \subset \mathbf{o}_{ij}$.

We make note of an abuse of notation in (3), as the group g in $F_{g,A}$ only corresponds to the same index of F_g in (2), if $B = A^c$, as can be seen in the example of Figure 1: If $B = A^c$, the three observations in $\mathbf{X}_{\mathcal{N}_A,A}$ are drawn from $F_{1,A}$, $F_{2,A}$ and $F_{3,A}$ respectively. However, if $B = \{2\}$, then observations two and three are now assumed to be drawn from a single distribution, which corresponds to a mixture of $F_{2,A}$ and $F_{3,A}$.

In short, the null hypothesis of (3) implies the null hypothesis of (2) because for $A = \mathbf{o}_{ij}$, observations coming from $F_{i,A}$ and $F_{j,A}$ are contained in $\mathbf{X}_{\mathcal{N}_A,A}$. Vice-versa, the null hypothesis of (2) implies the null hypothesis of (3) because A is nested in \mathbf{o}_{ij} for all F_i and F_j considered on A . This actually sketches the proof of the following result:

Proposition 1. Hypothesis testing problem (3) is equivalent to (2).

Tackling hypothesis testing problem (3) would be rather inefficient since we might test many times the same hypothesis when cycling through all $A \in \mathcal{A}$ and $B \in \mathcal{B}(A^c)$. However, the idea is that A and B will only be random draws from \mathcal{A} and $\mathcal{B}(A^c)$. This is discussed in the next section.

4. MCAR test Through Classification

In this section we introduce the classification-based statistic of our test and detail the implementation of our permutation approach, permuting the rows of the missingness matrix \mathbf{M} , to obtain a valid test.

4.1. Test Statistic U

Let us fix a projection $A \in \mathcal{A}$ and corresponding projection $B \in \mathcal{B}(A^c)$. We denote the induced collapsed class labels based on projections A and B by $Y^{(A,B)}$, by X_A the projection of the random vector X on A and correspondingly by x_A the projection on A of observation x in

$\mathbf{X}_{\mathcal{N}_{A,A}}$. Furthermore, we define for each $g \in G(A, B)$ and x in $\mathbf{X}_{\mathcal{N}_{A,A}}$ the following quantities:

$$\begin{aligned} p_g^{(A,B)}(x) &:= P(Y^{(A,B)} = g \mid X_A = x_A), \\ f_g^{(A,B)}(x) &:= P(x_A \mid Y^{(A,B)} = g), \\ \pi_g^{(A,B)} &:= P(Y^{(A,B)} = g). \end{aligned}$$

Let us fix $g \in G(A, B)$ as well. We reformulate the hypothesis testing problem (3):

$$\begin{aligned} H_{0,g}^{(A,B)} : f_g^{(A,B)} &= \frac{1}{1 - \pi_g^{(A,B)}} \sum_{j \in \{1, \dots, G(A,B)\} \setminus g} \pi_j^{(A,B)} f_j^{(A,B)} \\ \text{v.s.} & \\ H_{1,g}^{(A,B)} : f_g^{(A,B)} &\neq \frac{1}{1 - \pi_g^{(A,B)}} \sum_{j \in \{1, \dots, G(A,B)\} \setminus g} \pi_j^{(A,B)} f_j^{(A,B)}. \end{aligned} \tag{5}$$

Let $S_{f_g^{(A,B)}} \subset \mathcal{N}_A$ denote the indices of observations in $\mathbf{X}_{\mathcal{N}_{A,A}}$ that belong to class g . For each missingness pattern g , we now define the following statistic in analogy to Cai et al. (2020),

$$U_g^{(A,B)} := \frac{1}{|S_{f_g^{(A,B)}}|} \sum_{i \in S_{f_g^{(A,B)}}} \left(\log \frac{p_g^{(A,B)}(x_i)}{1 - p_g^{(A,B)}(x_i)} - \log \frac{\pi_g^{(A,B)}}{1 - \pi_g^{(A,B)}} \right). \tag{6}$$

This statistic is motivated by the following claim:

Lemma 1. The logarithm of the density ratio for testing (5) is given by $U_g^{(A,B)}$.

The main motivation for the form of this test-statistic is that one can use the same arguments as in Cai et al. (2020, Proposition 1) to show that a test based on $U_g^{(A,B)}$ will have the highest power among all tests for (5), according to the Neyman-Pearson Lemma. In addition, the test statistic converges to the Kullback-Leibler Divergence between $f_g^{(A,B)}$ and the mixture of the other densities, motivating the name of our MCAR test. A high value of KL-Divergence indicates that the distributions of two samples deviate strongly from each other.

Lemma 2. $U_g^{(A,B)}$ converges in probability to the Kullback-Leibler Divergence between $f_g^{(A,B)}$ and the mixture of the other densities:

$$U_g^{(A,B)} \rightarrow \mathbb{E}_{f_g} \left[\log \frac{f_g^{(A,B)}(X)(1 - \pi_g^{(A,B)})}{\sum_{j \in G(A,B) \setminus g} \pi_j^{(A,B)} f_j^{(A,B)}(X)} \right],$$

as n_g and $\sum_{j \in \{1, \dots, G\} \setminus g} n_j \rightarrow \infty$ and $n_g/n \rightarrow \pi_g^{(A,B)} \in (0, 1)$.

Since the statistic $U_g^{(A,B)}$ is evaluated only on cases $x \in S_{f_g^{(A,B)}}$, it holds that $f_g^{(A,B)}(x) = f_g^{*(A,B)}(x)$ and $p_g^{(A,B)}(x) = p_g^{*(A,B)}(x)$. This means that the projected complete and incomplete distributions coincide on the projected complete samples. Thus we are indeed asymptotically measuring the Kullback-Leibler Divergence between $f_g^{*(A,B)}$ and the mixture of the other densities.

Since there might be only very few observations for a single class g , we symmetrize the KL-Divergence. That is, we use the samples of all classes to evaluate the KL-Divergence and not only the samples of class g . Let $S_{f_{g^c(A,B)}} \subset \mathcal{N}_A$ denote the indices of observations in $\mathbf{X}_{\mathcal{N}_A, A}$ that belong to all other classes $G(A, B) \setminus g$. For each missingness pattern g , we will use, in the following, the difference between two of the above statistics, namely

$$U_g^{(A,B)} - U_{g^c}^{(A,B)} = \frac{1}{|S_{f_g^{(A,B)}}|} \sum_{i \in S_{f_g^{(A,B)}}} \log \frac{p_g^{(A,B)}(x_i)}{1 - p_g^{(A,B)}(x_i)} - \frac{1}{|S_{f_{g^c(A,B)}}|} \sum_{i \in S_{f_{g^c(A,B)}}} \log \frac{p_g^{(A,B)}(x_i)}{1 - p_g^{(A,B)}(x_i)}, \tag{7}$$

where the terms including the class probabilities $\pi_g^{(A,B)}$ cancel out. This difference converges to the symmetrized KL-Divergence between the mixture of $f_g^{(A,B)}$ and the remaining classes and is more sample efficient than only using $U_g^{(A,B)}$. The test statistic for fixed (A, B) is then given by

$$U^{(A,B)} := \sum_{g=1}^{G(A,B)} (U_g^{(A,B)} - U_{g^c}^{(A,B)}),$$

and the final test statistic is defined as

$$U := \mathbb{E}_{A \sim \kappa, B \sim \kappa(A^c)} [U^{(A,B)}]. \tag{8}$$

4.2. Practical Estimation of U

We estimate $p_g^{(A,B)}$ with a multiclass-classifier, yielding $\hat{p}_g^{(A,B)}$. Plugging-in this quantity into (7) yields $\hat{U}_g^{(A,B)} - \hat{U}_{g^c}^{(A,B)}$. We then estimate $U^{(A,B)}$ by

$$\hat{U}^{(A,B)} := \sum_{g=1}^{G(A,B)} (\hat{U}_g^{(A,B)} - \hat{U}_{g^c}^{(A,B)}).$$

Finally, we estimate U by

$$\hat{U} := \frac{1}{N} \sum_{i=1}^N \hat{U}^{(A_i, B_i)}, \quad (9)$$

where N is the number of draws of pairs of projections (A_i, B_i) , $i = 1, \dots, N$, with $A \in \mathcal{A}$ according to a distribution κ and $B \in \mathcal{B}(A^c)$ according to a distribution $\kappa(A^c)$.

Our chosen multiclass classifier is Random Forest (Breiman et al., 1984; Breiman, 2001), more specifically, the probability forest of Malley et al. (2011). That is, for each of the N projections, we fit a Random Forest with a specified number of trees, a parameter called `num.trees.proj`. Thus, for each tree (or group of trees) a random subset of variables and labels is chosen based on which the test statistic is computed. In each tree, we set `mtry` to the full dimension of the projection to not have an additional subsampling effect. This approach aligns naturally with the construction of Random Forest, as the overall approach might be seen as one aggregated Random Forest, which restricts the variables in each tree or group of trees to a random subset of variables. We finally use the OOB-samples for predicting $\hat{p}_g^{(A,B)}$.

The question remains how to sample the sets $(A_1, B_1), \dots, (A_N, B_N)$ at random. Our chosen approach is quite simple: we first randomly sample a number of dimensions r_1 by drawing uniformly from $\{1, \dots, p-1\}$. We then draw r_1 values without replacement from $\{1, \dots, p\}$ to obtain A . Similarly, we randomly draw a value r_2 from $\{1, \dots, p-r_1\}$ and then draw r_2 values without replacement from $\{1, \dots, p\} \setminus A$ to obtain B . We then consider $\mathbf{M}_{N,A,B}$, i.e., all patterns for the fully observed observations in A projected to B , and build the labels $Y^{(A,B)}$ based on the patterns in this matrix. This simple approach is used as a default, but one could also employ a more data-adaptive subsampling. In our algorithm, we might restrict the number of collapsed classes by selecting B corresponding to A accordingly. The parameter indicating the maximal number of collapsed classes allowed is given by `size.resp.set`. If set to 2, we reduce the multi-class problem to a two-class problem. In Algorithm 1 we provide the pseudo-code for the estimation of $\hat{U}^{(A,B)}$.

To ensure that the level is kept by a test based on the statistic \hat{U} for any choice of κ and $\kappa(A^c)$, we use a permutation approach, as detailed next.

Algorithm 1 $\text{Uhat}(\mathbf{X}, \mathbf{M}, A, B)$

Require: incomplete dataset \mathbf{X} , missingness indicator \mathbf{M} , projections A and B , number of trees per projection `num.trees.proj`, standard parameters of the Probability Forests, `size.resp.set`;

Recover the complete cases \mathcal{N}_A with respect to A ;

Generate the $|G(A, B)|$ collapsed class labels $Y^{(A,B)}$ from $\mathbf{M}_{\mathcal{N}_A, B}$;

Fit a multi-class probability forest with `num.trees.proj` trees and `mtry` full;

for $g \in G(A, B)$ **do**

Estimate $\hat{p}_g^{(A,B)}$ with the fitted forest above using out-of-bag probabilities;

Return the log-likelihood contribution $\hat{U}_g^{(A,B)} - \hat{U}_{g^c}^{(A,B)}$ for class g ;

end for

Average the log-likelihood ratio contributions $\hat{U}_g^{(A,B)} - \hat{U}_{g^c}^{(A,B)}$ from the $G(A, B)$ collapsed classes g to get the statistic $\hat{U}^{(A,B)}$;

return $\hat{U}^{(A,B)}$

4.3. Permutation Test

To ensure the correct level, we follow a permutation approach. Informally speaking, the permutation approach works in this context if the testing procedure can be replicated in exactly the same way on the randomly permuted class labels. This is not completely trivial in this case, as the labels are defined in each projection via the missingness matrix \mathbf{M} . It can be shown numerically that permuting the labels at the level of the projection does not conserve the level, as this is blind to the correspondence between the projections across the permutations.

The key to the correct permutation approach is to permute the rows of \mathbf{M} . That is, for L permutations σ_ℓ , $\ell = 1, \dots, L$, we obtain L matrices $\mathbf{M}_{\sigma_1}, \dots, \mathbf{M}_{\sigma_L}$ with only the rows permuted. Then we proceed as above: We sample $A \sim \kappa$, $B \sim \kappa(A^c)$ and for each permutation of rows σ_ℓ , $\ell = 1, \dots, L$, we calculate $U_{g, \sigma_\ell}^{(A,B)} - U_{g^c, \sigma_\ell}^{(A,B)}$ as in (7). Using $\hat{p}_g^{(A,B)}$ instead of $p_g^{(A,B)}$ this results in

$\hat{U}_{g,\sigma_\ell}^{(A,B)}$ and in the statistic

$$\hat{U}_{\sigma_\ell}^{(A,B)} := \sum_{g=1}^{G(A,B)} \hat{U}_{g,\sigma_\ell}^{(A,B)} - \hat{U}_{g^c,\sigma_\ell}^{(A,B)}.$$

We note that we do not need to refit the forest for this permutation approach to work. Instead, we can directly use $\hat{p}_g^{(A,B)}$ from the original Random Forest that we fitted on the original \mathbf{M} .

Finally, we calculate the empirical distribution of the test-statistic under the null, by calculating for $\ell = 1, \dots, L$,

$$\hat{U}_{\sigma_\ell} := \frac{1}{N} \sum_{j=1}^N \hat{U}_{\sigma_\ell}^{(A_j, B_j)}. \quad (10)$$

The p -value of the test is then obtained as usual by

$$Z := \frac{\sum_{\ell=1}^L \mathbb{1}\{\hat{U}_{\sigma_\ell} \geq \hat{U}\} + 1}{L + 1}. \quad (11)$$

Then it follows from standard theory on permutation tests that Z is a valid p -value:

Proposition 2. Under H_0 in (1), and Z as defined in (11), it holds for all $z \in [0, 1]$ that

$$\mathbb{P}(Z \leq z) \leq z. \quad (12)$$

Algorithm 2 summarizes the testing procedure.

Algorithm 2 PKLMtest(\mathbf{X})

Require: incomplete dataset \mathbf{X} , number of pairs of projections N , number of permutations L , number of trees per projection `num.trees.proj`, standard parameters of the Probability Forests, maximal number of collapsed classes `size.resp.set`;

Randomly permute the rows of \mathbf{M} L times to obtain $\mathbf{M}_{\sigma_1}, \dots, \mathbf{M}_{\sigma_L}$;

for $j = 1, \dots, N$ **do**

Sample a pair of projections (A_j, B_j) hierarchically according to $A_j \sim \kappa$ and $B_j \sim \kappa(A_j)$;

Calculate $\hat{U}^{(A_j, B_j)} = \text{Uhat}(\mathbf{X}, \mathbf{M}, A_j, B_j)$;

for $\ell = 1, \dots, L$ **do**

Calculate $\hat{U}_{\sigma_\ell}^{(A_j, B_j)} = \text{Uhat}(\mathbf{X}, \mathbf{M}_{\sigma_\ell}, A_j, B_j)$;

end for

end for

Average the statistics $\hat{U}^{(A_j, B_j)}, \hat{U}_{\sigma_\ell}^{(A_j, B_j)}$ over the couples of projections (A_j, B_j) to get the final statistic $\hat{U}, \hat{U}_{\sigma_\ell}, \ell = 1, \dots, L$;

Obtain the p -value with (11);

return p -value

5. Empirical Validation

In this section, we empirically showcase the power of our test in comparison to recent competitors on both simulated and real data. The simulation setting is set up along the lines of [Jamshidian and Jalal \(2010\)](#) and [Li and Yu \(2015\)](#) with a common MAR mechanism. For the real datasets we also add a random MAR generation through the function `ampute` of the R-package `mice`, see e.g., [Schouten et al. \(2018\)](#). As we did throughout the paper, we refer to our test as “PKLM”, the test of [Li and Yu \(2015\)](#) as “Q”, the test of [Little \(1988\)](#) as “Little” and finally the one of [Jamshidian and Jalal \(2010\)](#) as “JJ”. The Little-test is computed with the R-package `naniar` ([Tierney and Cook, 2023](#)), while the JJ-test uses the code of the R-package `MissMech` ([Jamshidian et al., 2014](#)). Finally, the code for the Q-test was kindly provided to us by the authors.

5.1. Simulated Data

We vary the sample size n , the number of dimensions p , and the number of complete observations, which we denote by r . Cases 1 – 8 describe the following different data distributions, similarly as in Li and Yu (2015) and in Jamshidian and Jalal (2010): Throughout, I_p is a covariance matrix with diagonal elements 1 and off-diagonal elements 0 while Σ is a covariance matrix with diagonal elements 1 and off-diagonal elements 0.7:

1. A standard multivariate normal distribution with mean 0 and covariance I_p ,
2. a correlated multivariate normal distribution with mean 0 and covariance Σ ,
3. a multivariate t -distribution with mean 0, covariance I_p and degree of freedom 4,
4. a correlated multivariate t -distribution with mean 0, covariance Σ and degree of freedom 4,
5. a multivariate uniform distribution which has independent uniform(0, 1) marginal distributions,
6. a correlated multivariate uniform distribution obtained by multiplying $\Sigma^{1/2}$ to the multivariate uniform distribution in 5,
7. a multivariate distribution obtained by generating $W = Z + 0.1Z^3$, where Z is from the standard multivariate normal distribution,
8. a multivariate Weibull distribution which has independent Weibull marginal distribution, and each Weibull marginal distribution has scale parameter 1 and shape parameter 2.

The above implements the fully observed \mathbf{X}^* . To compute the type-I error, we then simulate the MCAR mechanism where each value in the p columns of the missingness matrix \mathbf{M} has a probability of $1 - r^{1/p}$ being one and is otherwise zero. To compute the power, we simulate the MAR mechanism following the description in Li and Yu (2015): We generate \mathbf{M} such that the first column consists only of zeros so that the first variable is fully observed. Further, each value in the remaining $p - 1$ columns has a probability of $1 - r^{1/(p-1)}$ being one, while the rest is zero. This results, on average, in r rows in \mathbf{M} with only zeros, and thus in r fully observed rows in \mathbf{X} . Next, we sort the rows of \mathbf{M} into two groups, those that will be fully observed (complete group) and those that will have at least one missing value (missing group). So far, the generation is still MCAR. However now, for each row $i = 1, \dots, n$ we compare $\mathbf{X}_{i,1}^*$ with the mean of $\mathbf{X}_{\bullet,1}^*$, denoted

by \bar{X}_1 . If $\mathbf{X}_{i,1}^* < \bar{X}_1$, the corresponding row i is placed into the complete group with probability $1/6$, and with probability $5/6$ into the missing group. That is, with probability $1/6$, the row i is paired with a row in \mathbf{M} from the complete group, and with probability $5/6$, it is paired with a row from the missing group. Thus, in this case it is 5 times more likely that the row is placed in the missing group. On the other hand, if $\mathbf{X}_{i,1}^* \geq \bar{X}_1$ the situation reverses, and row i is 5 times more likely to be associated with a row in \mathbf{M} from the complete group. Assigning the rows of \mathbf{X}^* successively to the rows of \mathbf{M} like this results in \mathbf{X} with MAR missingness.

Each experiment was rerun `nsim = 300` times to compute type-I error and power. We used the following default hyperparameter setting for the computation of our PKLM-test: number of permutations `nrep = 30`, number of projections `num.proj = 100`, minimal node size in a tree `min.node.size = 10`, number of fitted trees per projection `num.trees.per.proj = 200` and maximal number of collapsed classes allowed in a projection `size.resp.set = 2`. We note that the choice of these hyperparameters is intriguingly simple: besides `size.resp.set`, it holds that “higher values are better”. Thus, as with RF in general, it is mostly a question of computational resources determining how large the values can be chosen. This is especially true for the number of trees for each forest, which should be relatively high in order to minimize additional randomness. We found `num.trees.per.proj = 200` to be a good compromise between speed and accuracy. As the level is guaranteed for any number of permutations, and we desired a choice of hyperparameters that would work for $p = 4$ as well as $p = 40$, we chose the number of permutations low (`nrep = 30`), but the number of projections relatively high (`num.proj = 100`). The only “difficult” parameter to set is `size.resp.set`, as there appears to be some loss in accuracy when the number of classes is larger than two. We thus found that `size.resp.set = 2`, generating two classes, works well in a wide range of examples.

As mentioned throughout the paper, the Q-test could not be calculated for a large range of settings.¹ In particular, computation times were infeasible for the setting $p = 10$ and $r = 0.1$, and for any configuration with $p = 20$ or $p = 40$. For the setting $n = 500$, $p = 10$ and $r = 0.1$ for instance, one test for case 2 took around 20 minutes to finish, implying an approximate overall

¹The largest number p reported in the paper of Li and Yu (2015) is 10, while r is at least 0.35.

computation time of $500 \cdot 8 \cdot 2 \cdot 20 = 16000$ minutes or approximately 110 24-hour days. This despite the fact that the R-code of the Q-test we received was well implemented. In the upcoming Tables 3 and 4 of results we always used the nominal level of $\alpha = 0.05$. We boldfaced the results for each row in the tables in the following manner: Whenever the type *I* error of a test is below or equal to 0.05 and the test has the best power, it will be boldfaced. If this is true for more than one test, they are all boldfaced. Additionally, we boldfaced all the type-*I* errors that are below or equal to the nominal level $\alpha = 0.05$ to indicate which tests holds the level on average in the given settings.

In the simulation set-up of $n = 200$ and $p = 4$, the Q-test is very powerful, while keeping the nominal level. The PKLM-test is rarely the most powerful here, however the power of the PKLM-test is often relatively close to the best power. As an example, in case 2 for $r = 0.65$, the Q-test has a power of 1 while the PKLM-test has a power 0.93, with both keeping the nominal level $\alpha = 0.05$.

In the set-up of $n = 500$ and $p = 10$, the overall picture changes. The PKLM-test is in all but two of the 24 cases the most powerful test, sometimes leaving the second-best test quite far behind. As an example, in case 3 for $r = 0.65$, the PKLM-test has a power of 0.85 while the Q- and the Little-test exhibit a power of 0.26 and 0.61, respectively. While the Little- and the JJ-test often show inflated levels, this is never a problem for the valid PKLM-test.

In the simulation set-up of $n = 500$ and $p = 20$, it appears as if the Little-test is a strong competitor. But this is only until one considers its type-I error. Though to a much lesser degree than the JJ-test, the type-I error is often heavily larger than the nominal level. Considering for instance case 4, the power of the Little-test is even slightly less than its actual type-I error for $r = 0.1$. In case 4 with $r = 0.35$, our test displays a power of 0.89 and keeps the level, while the Little-test only has a power of 0.33 despite having a grossly inflated type-I error. All of these problems are worsened for the JJ-test, which often displays an inflated type-I error in almost all cases and simulation set-ups. A similar story plays out in the case $r = 0.65$.

Finally, in the simulation set-up of $n = 1000$ and $p = 40$, the power of our test is again much better than that of all other tests. Interestingly, the PKLM-test tends to have higher power when the components of the distribution are not independent, such as in the cases 2, 4, 6, and 8. For

example, in case 1 for $r = 0.65$, PKLM has a power of 0.2, while for case 2 it has a power of 0.95. The main difference between these two cases is the strong positive correlation induced in case 2. This pattern repeats: in all correlated examples and for both $r = 0.65$ and $r = 0.35$, the PKLM has a power nearing 1, whereas in the independent versions, the power is closer to the type-I error. Thus, our test is able to use the dependencies in the data to its advantage, at least for $r = 0.65$ and $r = 0.35$, and can reach a very high power even for comparatively large p .

In summary, our test is very competitive even in small dimensions, where the Q-test is very powerful. It leaves behind all other tests by a wide margin as soon as one increases p . The Q-test remains strong in these situations as well, but becomes quickly infeasible as either p increases or the fraction of complete cases r decreases. Crucially, only the PKLM-test and the Q-test are able to consistently keep the nominal level over all experiments, with the Little- and JJ-test showing blatant inflation of the type-I error in many situations. This is the case despite the fact that simply checking the type-I error for a single level α (0.05 in this case) is far from sufficient to analyse the validity of a p -value.

As an illustration, we randomly chose one of the above experiments in which the Little-test kept the nominal level, e.g., in the simulation set up $n = 500$, $p = 10$, $r = 0.65$ in case 5. In Figure 2 we plot the empirical cumulative distribution functions (ecdf) of 500 p -values under the null (MCAR) of the four different tests. The red line is the $x = y$ line. In blue we plotted 100 ecdfs of a uniform(0, 1)-distribution. As described in Equation (12), a valid p -value has the property that the corresponding black ecdf values do not lie above the region defined by the blue lines. As Proposition 2 predicts, this is clearly the case for the PKLM-test. That the p -values appear rather discrete stems from the fact that we chose a low number of permutations ($nrep = 30$). The Q-test is sometimes overshooting the red line, though this appears to mostly stem from estimation error. In general, it is remarkable how closely the ecdfs of p -values from both the Q- and PKLM-test resemble the ecdf of a uniform sample. The JJ-test appears to consistently have $P(Z \leq z) \geq z$. The Little-test finally appears to produce a valid p -value as long as only values $z < 0.5$ are considered. For $z \geq 0.5$, the the ecdf clearly violates the requirement of a valid p -value. If there is no theoretical guarantee, it is thus important to not just check the type-I error at $\alpha = 0.05$, but to instead consider other levels, e.g., $\alpha = 0.1$.

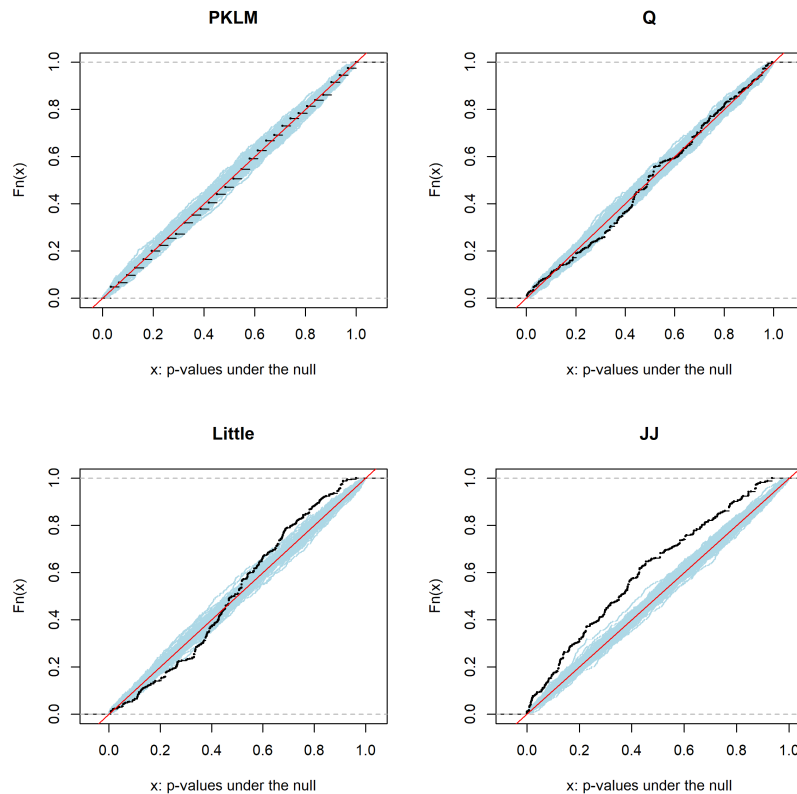


Figure 2: Example plot of cumulative distribution function values of the p -values under the null (MCAR) of the four different tests. The simulation set up is $n = 500$, $p = 10$, $r = 0.65$ in case 5, with 500 repetitions. The red line is the $x = y$ line, while the blue lines show 100 ecdfs of 500 simulated uniform random variables.

5.2. Real Data

We used 13 real datasets with varying number of observations n and dimensions p for further empirical assessment of the PKLM-test and comparison to the other three tests. The datasets are available in the UCI machine learning repository². We preprocessed the data by cancelling factor variables, in order to be able to run all other three tests. However, we kept numerical variables with only few unique values.

²<https://archive.ics.uci.edu/ml/index.php>

n	p	r	case	Power				Type-I Error			
				PKLM	Q	Little	JJ	PKLM	Q	Little	JJ
200	4	0.65	1	0.73	0.98	0.98	0.12	0.03	0.03	0.06	0.04
			2	0.93	1.00	0.96	0.04	0.03	0.06	0.06	0.05
			3	0.81	0.94	0.92	0.05	0.03	0.02	0.04	0.08
			4	0.89	0.97	0.91	0.05	0.01	0.03	0.05	0.05
			5	0.79	1.00	1.00	0.19	0.03	0.04	0.04	0.06
			6	0.90	1.00	0.99	0.20	0.03	0.04	0.03	0.13
			7	0.80	0.93	0.95	0.04	0.04	0.06	0.09	0.08
			8	0.72	0.92	0.90	0.26	0.03	0.05	0.04	0.08
200	4	0.35	1	0.79	0.98	0.97	0.04	0.03	0.04	0.04	0.13
			2	0.87	0.98	0.97	0.08	0.03	0.03	0.03	0.08
			3	0.82	0.97	0.90	0.16	0.03	0.03	0.06	0.12
			4	0.87	0.99	0.92	0.10	0.03	0.02	0.08	0.11
			5	0.79	0.99	0.99	0.10	0.04	0.05	0.05	0.08
			6	0.80	1.00	0.97	0.12	0.03	0.04	0.06	0.11
			7	0.79	0.98	0.92	0.09	0.03	0.05	0.07	0.06
			8	0.83	0.99	0.99	0.10	0.05	0.05	0.06	0.05
200	4	0.10	1	0.30	0.40	0.26	0.20	0.06	0.03	0.05	0.22
			2	0.35	0.50	0.27	0.12	0.03	0.10	0.05	0.18
			3	0.25	0.29	0.18	0.21	0.04	0.01	0.04	0.24
			4	0.37	0.42	0.17	0.19	0.03	0.03	0.03	0.17
			5	0.27	0.51	0.33	0.26	0.05	0.02	0.05	0.20
			6	0.31	0.40	0.27	0.24	0.03	0.03	0.04	0.17
			7	0.26	0.42	0.22	0.20	0.04	0.04	0.09	0.31
			8	0.31	0.39	0.32	0.23	0.03	0.03	0.04	0.18
500	10	0.65	1	0.93	0.89	0.88	0.09	0.05	0.06	0.06	0.05
			2	0.99	1.00	0.84	0.08	0.02	0.06	0.05	0.05
			3	0.85	0.26	0.61	0.12	0.02	0.05	0.18	0.10
			4	0.99	0.96	0.60	0.10	0.04	0.06	0.19	0.12
			5	0.89	0.98	0.96	0.16	0.04	0.05	0.03	0.10
			6	0.99	1.00	0.91	0.15	0.04	0.07	0.02	0.13
			7	0.90	0.61	0.68	0.09	0.02	0.07	0.12	0.07
			8	0.79	0.76	0.76	0.18	0.03	0.04	0.05	0.09
500	10	0.35	1	0.89	0.74	0.66	0.07	0.02	0.02	0.02	0.08
			2	0.99	0.99	0.69	0.09	0.03	0.06	0.03	0.11
			3	0.88	0.33	0.51	0.14	0.04	0.05	0.18	0.11
			4	0.98	0.91	0.48	0.12	0.04	0.08	0.20	0.10
			5	0.91	0.92	0.83	0.12	0.04	0.06	0.04	0.12
			6	0.98	1.00	0.75	0.09	0.03	0.08	0.04	0.11
			7	0.89	0.46	0.52	0.05	0.03	0.03	0.08	0.11
			8	0.92	0.78	0.74	0.10	0.05	0.06	0.06	0.07
500	10	0.10	1	0.31	–	0.06	0.12	0.02	–	0.03	0.10
			2	0.45	–	0.07	0.12	0.03	–	0.03	0.07
			3	0.34	–	0.18	0.16	0.03	–	0.19	0.14
			4	0.45	–	0.20	0.16	0.02	–	0.22	0.11
			5	0.33	–	0.04	0.12	0.06	–	0.02	0.14
			6	0.45	–	0.03	0.08	0.05	–	0.01	0.12
			7	0.34	–	0.12	0.09	0.05	–	0.09	0.15
			8	0.34	–	0.04	0.16	0.03	–	0.05	0.13

Table 3: Simulated power and type-I error of PKLM, Q, Little and JJ for $n = 200$, $p = 4$ and $n = 500$, $p = 10$. We use $r = 0.65$, 0.35 and 0.1 . Cases 1 – 8 describe different data distributions. The experiments were repeated 300 times and the parameter setting for PKLM described above was used.

n	p	r	case	Power				Type-I Error			
				PKLM	Q	Little	JJ	PKLM	Q	Little	JJ
500	20	0.65	1	0.39	–	0.36	0.06	0.02	–	0.05	0.09
			2	0.91	–	0.48	0.08	0.03	–	0.05	0.10
			3	0.33	–	0.49	0.20	0.03	–	0.24	0.11
			4	0.90	–	0.40	0.14	0.04	–	0.22	0.11
			5	0.32	–	0.64	0.14	0.04	–	0.04	0.08
			6	0.93	–	0.39	0.25	0.04	–	0.01	0.09
			7	0.33	–	0.37	0.07	0.03	–	0.09	0.10
			8	0.23	–	0.25	0.14	0.04	–	0.06	0.03
500	20	0.35	1	0.45	–	0.22	0.08	0.03	–	0.04	0.09
			2	0.90	–	0.22	0.09	0.03	–	0.04	0.08
			3	0.43	–	0.35	0.18	0.02	–	0.34	0.12
			4	0.89	–	0.33	0.20	0.03	–	0.31	0.15
			5	0.46	–	0.24	0.09	0.02	–	0.02	0.12
			6	0.91	–	0.14	0.14	0.02	–	0.03	0.10
			7	0.41	–	0.22	0.11	0.02	–	0.11	0.10
			8	0.52	–	0.18	0.08	0.03	–	0.04	0.07
500	20	0.10	1	0.13	–	0.00	0.14	0.03	–	0.00	0.10
			2	0.24	–	0.01	0.14	0.04	–	0.01	0.12
			3	0.08	–	0.21	0.16	0.06	–	0.22	0.10
			4	0.26	–	0.27	0.08	0.04	–	0.31	0.13
			5	0.12	–	0.00	0.10	0.03	–	0.00	0.19
			6	0.19	–	0.00	0.11	0.05	–	0.00	0.18
			7	0.07	–	0.08	0.12	0.04	–	0.07	0.12
			8	0.07	–	0.02	0.11	0.04	–	0.00	0.16
1000	40	0.65	1	0.20	–	0.00	0.09	0.05	–	0.00	0.15
			2	0.95	–	0.00	0.12	0.03	–	0.00	0.14
			3	0.23	–	0.00	0.29	0.02	–	0.00	0.17
			4	0.94	–	0.00	0.26	0.05	–	0.00	0.17
			5	0.16	–	0.00	0.30	0.02	–	0.00	0.19
			6	0.97	–	0.00	0.26	0.02	–	0.00	0.19
			7	0.23	–	0.00	0.11	0.02	–	0.00	0.10
			8	0.13	–	0.00	0.17	0.03	–	0.00	0.12
1000	40	0.35	1	0.35	–	0.00	0.12	0.02	–	0.00	0.11
			2	0.97	–	0.00	0.13	0.05	–	0.00	0.10
			3	0.37	–	0.00	0.30	0.03	–	0.00	0.30
			4	0.96	–	0.00	0.33	0.04	–	0.00	0.27
			5	0.32	–	0.00	0.14	0.04	–	0.00	0.11
			6	0.98	–	0.00	0.16	0.03	–	0.00	0.10
			7	0.36	–	0.00	0.11	0.02	–	0.00	0.08
			8	0.30	–	0.00	0.16	0.02	–	0.00	0.10
1000	40	0.10	1	0.08	–	0.00	0.15	0.02	–	0.00	0.12
			2	0.32	–	0.00	0.12	0.02	–	0.00	0.10
			3	0.06	–	0.00	0.13	0.05	–	0.00	0.20
			4	0.25	–	0.00	0.25	0.03	–	0.00	0.28
			5	0.08	–	0.00	0.11	0.03	–	0.00	0.09
			6	0.27	–	0.00	0.09	0.04	–	0.00	0.11
			7	0.07	–	0.00	0.16	0.03	–	0.00	0.13
			8	0.07	–	0.00	0.15	0.04	–	0.00	0.08

Table 4: Simulated power and type-I error of PKLM, Q, Little and JJ for $n = 500$, $p = 20$ and $n = 1000$, $p = 40$. We use $r = 0.65$, 0.35 and 0.1 . Cases 1 – 8 describe different data distributions. The experiments were repeated 300 times and the parameter setting for PKLM described above was used.

dataset	n	p	Power				Type-I Error			
			PKLM	Q	Little	JJ	PKLM	Q	Little	JJ
iris	150	4	0.41	0.91	0.84	0.27	0.03	0.04	0.03	0.16
blood.transfusion	748	4	0.48	0.97	1.00	NA	0.01	0.06	0.04	NA
airfoil	1503	6	0.92	0.13	0.17	0.09	0.02	0.03	0.06	0.42
seeds	210	7	0.64	0.74	0.57	0.24	0.05	0.02	0.02	0.10
yacht	308	7	0.60	0.56	0.76	0.24	0.03	0.07	0.05	0.24
yeast	1484	8	0.82	0.52	0.15	0.14	0.05	0.06	0.23	0.85
glass	214	9	0.10	0.02	0.20	0.20	0.01	0.00	0.03	0.33
concrete.compression	1030	9	0.64	0.48	0.81	0.47	0.04	0.04	0.05	0.41
wine.quality.red	1599	11	0.81	–	0.72	0.80	0.04	–	0.15	0.52
wine.quality.white	4898	11	0.98	–	0.96	0.87	0.04	–	0.10	0.79
planning.relax	182	12	0.29	–	0.20	0.14	0.00	–	0.00	NA
climate.model.crashes	540	19	0.18	–	0.22	0.47	0.00	–	0.00	NA
ionosphere	351	32	0.45	–	0.97	0.18	0.00	–	0.06	NA

Table 5: Simulated power and level of PKLM, Q, Little and JJ for 13 real datasets. We use $p_{miss} = 0.3$. The experiments were repeated 300 times and the parameter setting for PKLM described above was used. The NAs for some values of the JJ-test indicate that the test was not computable in any of the 300 repetitions due to not enough observations in enough usable missingness groups.

For the generation of the NAs, we use an overall probability of missingness of $p_{miss} = 0.3$ (not to be confused with r from the last subsection, denoting the number of complete cases). We used a random MAR generation through the function `ampute` of the R-package `mice`. This function can randomly generate realistic MAR mechanisms, see e.g., Schouten et al. (2018). Each experiment was run `nsim = 300` times to compute the type-I error and power. We used the following hyperparameter setting for the computation of our PKLM-test: number of permutations `nrep = 30`, number of projections `num.proj = 300`, minimal node size in a tree `min.node.size = 10`, number of fitted trees per projection `num.trees.per.proj = 200` and maximal number of collapsed classes allowed in a projection `size.resp.set = 2`. The results are shown in Table 5. Our test is again very competitive with the best power in 7 out of 13 datasets, conditional on valid type-I errors. The Little-test shows also often good performance, though given the problematic level displayed in the previous section, this has to be considered with some

care. The Q-test also has relatively high power in the situations where it can be calculated. However, due to computational time we only run the Q-test for $p \leq 10$. All in all, we see that the Q-test quickly gets infeasible for large p and n and the advantage of the PKLM-test strengthens with increasing p .

6. Extension

In addition to the “global test” of MCAR, we can study the effect of single variables: For any given variable $k = 1, \dots, p$, we can calculate

$$\hat{U}^{-k} = \frac{1}{|\mathcal{P}_{-k}|} \sum_{i \in \mathcal{P}_{-k}} \hat{U}^{(A_i, B_i)},$$

where \mathcal{P}_{-k} are all pairs of projections (A_i, B_i) from the N randomly chosen ones, with B_i not containing variable k . We can use the analogous calculation based on the permuted missingness matrix \mathbf{M}

$$\hat{U}_{\sigma_\ell}^{-k} = \frac{1}{|\mathcal{P}_{-k}|} \sum_{j \in \mathcal{P}_{-k}} \hat{U}_{\sigma_\ell}^{(A_j, B_j)},$$

to obtain the p -value as in (11). This “partial” p -value is valid and corresponds to the effect of removing the patterns induced by variable k . Indeed, assume the difference in the distribution of two patterns stems from a variable j alone. If $j \in B$, a perfect classifier will be able to reliably differentiate the two, leading to a high value for \hat{U}^{-k} relative to the permutation values. If j is not forming the labels, we will not test these two classes against each other and thus not be able to spot this difference. As such, we might expect to see a high p -value for \hat{U}^{-j} , when variable j is removed, but a tendency to low p -values for \hat{U}^{-k} , $k \neq j$.

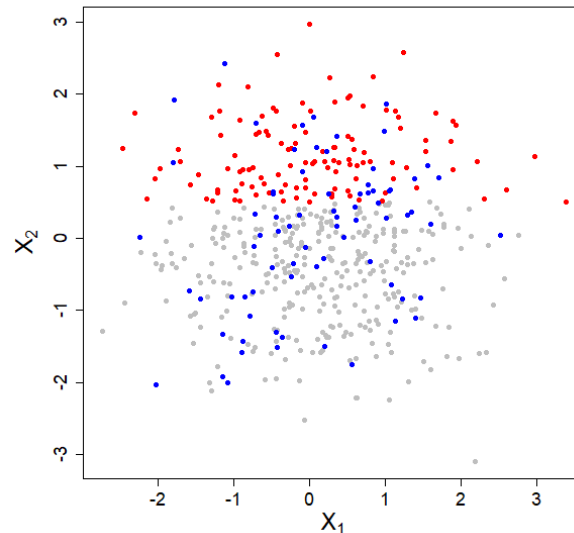


Figure 3: X_1 and X_2 of the fully observed data in the simulated example of Section 6. In red: Points with missing values in X_1 , in blue: points with missing values in X_2 . The blue points are randomly scattered, independently of the value of X_1 , while in the red points, there is a visible trend towards having more missing values in X_1 for higher values of variable X_2 .

We illustrate the usefulness of partial p -values with an example. Let $C_{-k} = \{1, \dots, p\} \setminus \{k\}$. We assume $\mathbf{X}_{\bullet, C_{-k}}$ has a MCAR missingness structure, in particular, we simulate below the MCAR mechanism described in Section 5.1 with $r = 0.65$. Let $k = 1$ and assume that this first column of observations $\mathbf{X}_{\bullet, 1}$ has missingness depending on the observed values of $\mathbf{X}_{\bullet, 2}$. For instance, each value is missing if the mean of the corresponding row $\mathbf{X}_{j, 2}$ is larger than 0.5. In this simple example \mathbf{X} is MAR, but $\mathbf{X}_{\bullet, C_{-1}}$ is MCAR. We simulate this example, with $p = 4$ and $n = 500$, $\mathbf{X}_{i, \bullet}$ being independent standard Gaussian and the MAR/MCAR mechanism as described above. The first two fully observed components, X_1 and X_2 , are shown in Figure 3. As before, we set `num.trees.proj=200` and use 100 projections. In this example, we are only able to spot any difference when $j = 1$ is used to build the labels.

Our test reliably delivers small p -values (≤ 0.05) for the three partial tests based on projections potentially including variable 1, i.e., sets of projections \mathcal{P}_{-2} , \mathcal{P}_{-3} , and \mathcal{P}_{-4} and a high

p -value for the partial test based on \mathcal{P}_{-1} . Thus in this sense, the test detects that the main culprit of the MAR mechanism lies in the first variable.

7. Concluding Remarks

In this paper we presented the powerful, flexible and easy-to-use PKLM-test for the MCAR assumption on the missingness mechanism of a dataset. We proved the validity of the p -value of the test and showed its power over a wide range of distributions. We also provided an extension allowing to do partial tests, that may shed light on the source of the violation of the MCAR assumption. Naturally, with some slight adaptations the test can be used as a general test of homogeneity of G different groups in the sense that it tests whether G different groups have the same distribution.

Appendix

A. Proofs

Proposition 1. Hypothesis testing problem (3) is equivalent to (2).

Proof. We first show H_0 of (2) implies H_0 of (3). Let A, B be arbitrary. If they are such that there is only one label, there is nothing to test, so we may assume to have $|G(A, B)| \geq 2$ patterns in $\mathbf{X}_{\mathcal{N}_A, A}$. This means that $A \subset \mathbf{o}_{ij}$ for all patterns $i, j \in G(A, B)$. This simply follows because, by construction, each of the $|G(A, B)|$ patterns in $\mathbf{X}_{\mathcal{N}_A, A}$ has the elements in A fully observed. But since by assumption for all $i, j \in \{1, \dots, G\}$, $F_{i, \mathbf{o}_{ij}} = F_{j, \mathbf{o}_{ij}}$ and $A \subset \mathbf{o}_{ij}$, this immediately implies that $F_{i, A} = F_{j, A}$ for all $i, j \in \{1, \dots, G\}$ and thus $F_{g, A} = \sum_{j \in G(A, B) \setminus g} \omega_j^g F_{j, A}$. Since A, B were arbitrary, one direction follows.

We now show that H_0 of (3) implies H_0 of (2). The proof is based on the following claim: Consider G arbitrary distribution functions F_1, \dots, F_G and weights $(\omega_j^g)_{j=1}^{G-1}$, $j = 1, \dots, G$ such that $\sum_{j=1}^{G-1} \omega_j^g = 1$ for all j . Then

$$F_g = \sum_{j \in \{1, \dots, G\} \setminus g} \omega_j^g F_j, \forall g \in \{1, \dots, G\} \implies F_i = F_j, \forall i \neq j \in \{1, \dots, G\}. \tag{13}$$

We prove the implication by induction: Consider first $G = 3$. Assuming the LHS of (13) and plugging the equation for F_2 into the equation for F_1 , we obtain:

$$\begin{aligned} F_1 &= w_2^1 w_1^2 F_1 + w_2^1 w_3^2 F_3 + w_3^1 F_3 \\ &= w_2^1 w_1^2 F_1 + (w_2^1 w_3^2 + w_3^1) F_3, \end{aligned}$$

which implies $(1 - w_1^2 w_2^1) F_1 = (w_2^1 w_3^2 + w_3^1) F_3$. Since

$$1 = w_2^1 + w_3^1 = w_2^1 (w_3^2 + w_1^2) + w_3^1 = w_2^1 w_3^2 + w_2^1 w_1^2 + w_3^1,$$

we have the equality $(1 - w_1^2 w_2^1) = (w_2^1 w_3^2 + w_3^1)$ and thus $F_1 = F_3$. Plugging this back into the equivalent equation for F_2 , we obtain $F_1 = F_2 = F_3$. Now assume (13) is true for G distributions F_1, \dots, F_G and we now would like to prove it for $G + 1$. Assume wlog that the weight of F_2 in the equation of F_1 is nonzero (there will always be at least one such distribution F_2, \dots, F_G). Using

the same trick as above, we may plug say the equation for F_2 into F_1 , thereby reducing the number of equations/distributions to G . By the induction assumption this implies that $F_1 = F_3 = \dots = F_G$. But immediately this also implies that $F_2 = F_1$ and implies (13). With this result we can now proof the that H_0 of (3) implies H_0 of (2).

Take two arbitrary groups i, j and $A = \mathbf{o}_{ij}$ and take $B = A^c$. To ease notation we just wlog take $i = 1$ and $j = 2$. Then $A = \mathbf{o}_{12}$ contains the dimensions for which patterns 1 and 2 have fully observed values. Thus, observations in $\mathbf{X}_{\mathcal{N}_A, A}$ contain draws from $F_{1, \mathbf{o}_{12}}$ and $F_{2, \mathbf{o}_{12}}$. Since by assumption

$$H_0 : F_{g,A} = \sum_{j \in G(A,B) \setminus g} \omega_j^g F_{j,A}, \forall g \in G(A,B), \tag{14}$$

it follows by (13), that $F_{i,A} = F_{j,A}$ for all $i, j \in G(A,B)$ and thus in particular, $F_{1,A} = F_{2,A}$. Since we will have $A = \mathbf{o}_{ij}$ for all groups $i \neq j$, H_0 of (2) holds.

Lemma 1. The logarithm of the density ratio for testing (5) is given by $U_g^{(A,B)}$.

Proof. Based on the definitions of $p_g^{(A,B)}(x)$, $f_g^{(A,B)}(x)$ and $\pi_g^{(A,B)}$ we obtain by Bayes Rule,

$$p_g^{(A,B)}(x) = \frac{f_g^{(A,B)}(x)\pi_g^{(A,B)}}{\sum_{j \in G(A,B)} \pi_j^{(A,B)} f_j^{(A,B)}(x)}, \tag{15}$$

assuming the existence of densities f_g of distributions F_g for each $g \in G(A,B)$. Following the same steps as in Cai et al. (2020), we get that the logarithm of the (joint) density ratio for testing H_0 vs H_1 of (5), given by

$$\log \frac{f_g^{(A,B)}(x)(1 - \pi_g^{(A,B)})}{\sum_{j \in G(A,B) \setminus g} \pi_j^{(A,B)} f_j^{(A,B)}(x)}. \tag{16}$$

We reformulate the fraction in (16) in terms of $p_g^{(A,B)}$, starting from (15):

$$\begin{aligned} p_g^{(A,B)}(x) \sum_{j \in G(A,B) \setminus g} \pi_j^{(A,B)} f_j^{(A,B)}(x) &= (\pi_g^{(A,B)} - p_g^{(A,B)}(x)\pi_g^{(A,B)})f_g^{(A,B)}(x) \\ &= \pi_g^{(A,B)}(1 - p_g^{(A,B)}(x))f_g^{(A,B)}(x). \end{aligned}$$

Thus, the inside of the logarithm of (16) is given by the following function of $p_g^{(A,B)}$:

$$\frac{f_g^{(A,B)}(x)(1 - \pi_g^{(A,B)})}{\sum_{j \in G(A,B) \setminus g} \pi_j^{(A,B)} f_j^{(A,B)}(x)} = \frac{1 - \pi_g^{(A,B)}}{\pi_g^{(A,B)}} \frac{p_g^{(A,B)}(x)}{1 - p_g^{(A,B)}(x)}.$$

Lemma 2. $U_g^{(A,B)}$ converges in probability to the Kullback-Leibler Divergence between $f_g^{(A,B)}$ and the mixture of the other densities:

$$U_g^{(A,B)} \rightarrow \mathbb{E}_{f_g} \left[\log \frac{f_g^{(A,B)}(X)(1 - \pi_g^{(A,B)})}{\sum_{j \in G(A,B) \setminus g} \pi_j^{(A,B)} f_j^{(A,B)}(X)} \right],$$

as n_g and $\sum_{j \in \{1, \dots, G\} \setminus g} n_j \rightarrow \infty$ and $n_g/n \rightarrow \pi_g^{(A,B)} \in (0, 1)$.

Proof. From the proof of Lemma 1, we know that $U_g^{(A,B)}$ can be rewritten as

$$\begin{aligned} U_g^{(A,B)} &:= \frac{1}{|S_{f_g^{(A,B)}}|} \sum_{i \in S_{f_g^{(A,B)}}} \left(\log \frac{p_g^{(A,B)}(x_i)}{1 - p_g^{(A,B)}(x_i)} - \log \frac{\pi_g^{(A,B)}}{1 - \pi_g^{(A,B)}} \right) \\ &= \frac{1}{n_g} \sum_{i \in S_{f_g^{(A,B)}}} \log \frac{f_g^{(A,B)}(x_i)(1 - \pi_g^{(A,B)})}{\sum_{j \in G(A,B) \setminus g} \pi_j^{(A,B)} f_j^{(A,B)}(x_i)}. \end{aligned} \tag{17}$$

Since $n_g/n \rightarrow \pi_g^{(A,B)} \in (0, 1)$ and the x_i are i.i.d., the result follows from the law of large numbers.

Proposition 2. Under H_0 in (1), and Z as defined in (11), it holds for all $z \in [0, 1]$ that

$$\mathbb{P}(Z \leq z) \leq z. \tag{12}$$

Proof. Let $\mathbf{A} = (A_1, \dots, A_N)$ and $\mathbf{B} = (B_1, \dots, B_N)$ be two sets of N projections. Let G_1, \dots, G_{L^*} be all possible permutations of the rows of the missingness matrix \mathbf{M} , such that

$$G_\ell(\mathbf{X}^*, \mathbf{M}, \mathbf{A}, \mathbf{B}) = (\mathbf{X}^*, \mathbf{M}_{\sigma_\ell}, \mathbf{A}, \mathbf{B}),$$

for $\ell = 1, \dots, L^*$. Note that, since we are only considering fully observed observations for all projections in \mathbf{A} , \hat{U} , a function of $(\mathbf{X}, \mathbf{M}, \mathbf{A}, \mathbf{B})$, is indeed a function of $(\mathbf{X}^*, \mathbf{M}, \mathbf{A}, \mathbf{B})$, while \hat{U}_{σ_ℓ} is a function of $G_\ell(\mathbf{X}^*, \mathbf{M}, \mathbf{A}, \mathbf{B})$. It also holds that under the null, that is under MCAR, that

$$(\mathbf{X}^*, \mathbf{M}, \mathbf{A}, \mathbf{B}) \stackrel{D}{=} (\mathbf{X}^*, \mathbf{M}_{\sigma_\ell}, \mathbf{A}, \mathbf{B}) = G_\ell(\mathbf{X}^*, \mathbf{M}, \mathbf{A}, \mathbf{B}) \quad \forall \ell = 1, \dots, L^*. \tag{18}$$

This is true because, under MCAR, \mathbf{M} and \mathbf{X}^* are independent. Since by the i.i.d. assumption also $\mathbf{M}_{\sigma_\ell} \stackrel{D}{=} \mathbf{M}$ for all $\ell = 1, \dots, L^*$ and since \mathbf{A} , \mathbf{B} are also independent of \mathbf{M} , (18) follows. As outlined for example in Hemerik and Goeman (2018), this implies that under H_0 ,

$$\mathbb{P}(Z \leq z \mid \mathbf{A}, \mathbf{B}) \leq z.$$

Integrating over (\mathbf{A}, \mathbf{B}) , results in (12).

B. Additional Details and Computation Times

Here we provide more implementation details, discuss the complexity calculations in Table 1 and show computation times of the different tests in the experiments.

Numerical truncation. In order to avoid numerical issues when calculating the density ratio with Expression (6) or the log thereof, if we get predicted probabilities \hat{p}_A close to 0 or 1, we apply the following truncation function to \hat{p}_A :

$$p(x) = \min(\max(x, 10^{-9}), 1 - 10^{-9}).$$

Hyperparameter Selection. Generally speaking, it holds that “the more the better”, certainly for the parameters N , L and `num.trees.proj`. As such, the choice of those three parameters depends mostly on the computational power available to the user. For `size.resp.set`, this is not quite as clear, though we found a value of two to work well in most situations.

PLKM Test. We first consider the complexity of one Random Forest, which is in this case

$$\text{num.tree} \cdot pn \log(n).$$

Note that this includes the calculation of \hat{p} on the test sample through the OOB-error. In total we do this `num.proj` times. However, we consider `num.tree` and `num.proj` independent of n and p and thus treating it as constant. In this case we end up with $pn \log(n)$. Finally we need to calculate the statistics U and repeat this number of calculations a fixed number of times. This would add a factor Bn , where again we assume that B does not grow with n and p . As this is negligible compared to $pn \log(n)$, the complexity is given as $\mathcal{O}(pn \log(n))$.

Q-test. The Q-test compares all groups leading to a complexity of G^2 to compare each group with any other. Additionally, the statistic used is an MMD type, so the complexity is $(n_1 + n_2)^2$, where n_1, n_2 are the respective group sizes. The group size can be at worst n/G , which together results in $\mathcal{O}(n^2)$. The bootstrap on the other hand can also be ignored, as it simply results in a constant factor multiplied to n^2 .

JJ and Little-test. Both JJ- and Little-test rely on covariance estimation which scales as np^2 . This gives the $\mathcal{O}(np^2)$ complexity for the Little-test. For the JJ-test one also needs an ordering operation to obtain the test statistics, with complexity $n \log(n)$, which results in overall complexity $\mathcal{O}(n(p^2 + \log(n)))$.

As mentioned above, Table 1 just shows how the complexity scales in n and p and, in case of our test, treats the number of projections as constants. One might argue that the number of projections should be a function of p as well. Similarly, for “small” p and small number of groups G , the Q-test can be faster than ours. Still the complexities provide a good illustration of how quickly the Q-test can become infeasible, when the number of groups (often a function of p) and/or the number of observation increases.

C. Example of Yuan et al. (2018)

Yuan et al. (2018) study settings where group means and variances are approximately equal across missingness patterns, such that MCAR tests based on differences in means and variances, such as the Little-test, have no power. We study one such example here: Let $p = 2$ and (Z_1, Z_2) be jointly multivariate normal with correlation zero and let $X_1 = Z_1$ and

$$X_2 = 0.5Z_1 + (1 - 0.25)^{1/2}Z_2.$$

We set X_2 to NA if

$$X_1 \in (-\infty, -1.932] \cup (-0.314, 0.314] \cup (1.932, \infty).$$

This corresponds to around 30% missing values. Figure 4 displays a histogram, plotting all observations of X_1 with X_2 missing for a simulation of $n = 10'000$. This corresponds to the MAR example used in Yuan et al. (2018, Section 3) and we refer to their paper for more details.

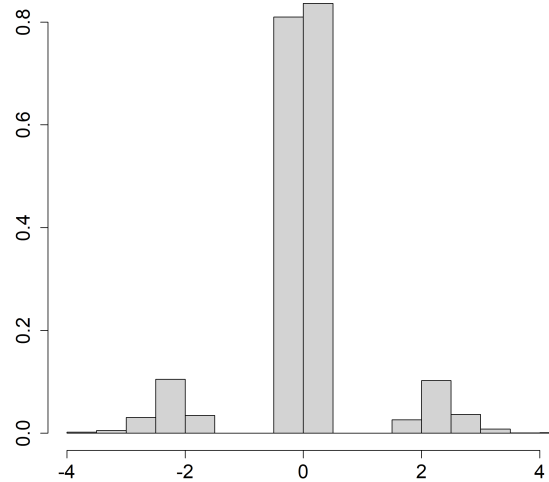


Figure 4: Histogram with relative frequencies of X_1 if the corresponding X_2 is NA.

We simulate the above distribution for $n = 1000$ and run our PKLM-test with the same parameters as described in Section 5.1. Though the deviation from MCAR cannot be detected through the first two moments in this example, our test reliably reaches a power of 1.

References

- Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1):5–32.
- Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). Classification and regression trees.
- Brown, S. M., Doom, J. R., Lechuga-Peña, S., Watamura, S. E., and Koppels, T. (2020). Stress and parenting during the global covid-19 pandemic. *Child Abuse & Neglect*, 110:104699.
- Buuren, S. and Groothuis-Oudshoorn, C. (2011). Mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, 45.
- Cai, H., Goggin, B., and Jiang, Q. (2020). Two-sample test based on classification probability. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 13(1):5–13.
- Charles, N. E., Strong, S. J., Burns, L. C., Bullerjahn, M. R., and Serafine, K. M. (2021). Increased mood disorder symptoms, perceived stress, and alcohol use among college students during the covid-19 pandemic. *Psychiatry Research*, 296:113706.
- Chen, H. and Little, R. (1999). A test of missing completely at random for generalised estimating equations with missing data. *Biometrika*, 86(1):1–13.
- de Vos, J. A., Radstaak, M., Bohlmeijer, E. T., and Westerhof, G. J. (2022). Modelling trajectories of change in psychopathology and well-being during eating disorder outpatient treatment. *Psychotherapy Research*, 0(0):1–13.
- Deng, Y., Chang, C., Ido, M. S., and Long, Q. (2016). Multiple imputation for general missing data patterns in the presence of high-dimensional data. *Scientific reports*, 6:21689–21689.
- Hawes, M. T., Szency, A. K., Klein, D. N., Hajcak, G., and Nelson, B. D. (2021). Increases in depression and anxiety symptoms in adolescents and young adults during the COVID-19 pandemic. *Psychological Medicine*, page 1–9.
- Hediger, S., Michel, L., and Näf, J. (2022). On the use of random forest for two-sample testing. *Computational Statistics & Data Analysis*, 170:107435.

- Hemerik, J. and Goeman, J. (2018). Exact testing with random permutations. *TEST*, 27(4):811–825.
- Jamshidian, M. and Jalal, S. (2010). Tests of homoscedasticity, normality, and missing completely at random for incomplete multivariate data. *Psychometrika*, 75(4):649–674.
- Jamshidian, M., Jalal, S., and Jansen, C. (2014). MissMech: An R package for testing homoscedasticity, multivariate normality, and missing completely at random (mcar). *Journal of Statistical Software*, 56(6):1–31.
- Kim, I., Ramdas, A., Singh, A., and Wasserman, L. (2021). Classification accuracy as a proxy for two-sample testing. *The Annals of Statistics*, 49(1):411 – 434.
- Kim, K. H. and Bentler, P. M. (2002). Tests of homogeneity of means and covariance matrices for multivariate incomplete data. *Psychometrika*, 67(4):609–623.
- Langer, Á. I., Ponce, F. P., Ordóñez-Carrasco, J. L., et al. (2024). Psychometric evidence of the Acceptance and Action Questionnaire-II (AAQ-II): an item response theory analysis in university students from Chile. *BMC Psychology*, 12:111.
- Lehmann, E. L. and Romano, J. P. (2005). *Testing Statistical Hypotheses*. Springer Texts in Statistics. Springer, third edition.
- Li, J. and Yu, Y. (2015). A nonparametric test of missing completely at random for incomplete multivariate data. *Psychometrika*, 80(3):707–726.
- 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.
- Little, R. J. A. and Rubin, D. B. (1986). *Statistical Analysis with Missing Data*. John Wiley & Sons, Inc.
- Malley, J., Kruppa, J., Dasgupta, A., Malley, K., and Ziegler, A. (2011). Probability machines consistent probability estimation using nonparametric learning machines. *Methods of information in medicine*, 51:74–81.

- Muzellec, B., Josse, J., Boyer, C., and Cuturi, M. (2020). Missing data imputation using Optimal Transport. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pages 7130–7140. PMLR.
- Näf, J., Spohn, M.-L., Michel, L., and Meinshausen, N. (2023). Imputation scores. *The Annals of Applied Statistics*, 17(3):2452 – 2472.
- Rajeb, M., Wang, Y., Man, K., et al. (2023). Students' acceptance of online learning in developing nations: scale development and validation. *Education Tech Research Dev*, 71:767–792.
- Rhoads, C. H. (2012). Problems with tests of the missingness mechanism in quantitative policy studies. *Statistics, Politics, and Policy*, 3(1).
- Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3):581–592.
- Schouten, R. M., Lugtig, P., and Vink, G. (2018). Generating missing values for simulation purposes: a multivariate amputation procedure. *Journal of Statistical Computation and Simulation*, 88(15):2909–2930.
- Sun, X. and Chen, M. (2022). Associations between perceived material deprivation, social support and violent victimization among chinese children. *Child Abuse & Neglect*, 127:105583.
- Tierney, N. and Cook, D. (2023). Expanding tidy data principles to facilitate missing data exploration, visualization and assessment of imputations. *Journal of Statistical Software*, 105(7):1–31.
- Yuan, K.-H., Jamshidian, M., and Kano, Y. (2018). Missing data mechanisms and homogeneity of means and variances–covariances. *Psychometrika*, 83(2):425–442.
- Zarate, D., Hobson, B. A., March, E., Griffiths, M. D., and Stavropoulos, V. (2023). Psychometric properties of the bergen social media addiction scale: An analysis using item response theory. *Addictive Behaviors Reports*, 17:100473.
- Zhang, S., Han, P., and Wu, C. (2019). A unified empirical likelihood approach for testing mcar and subsequent estimation. *Scandinavian Journal of Statistics*, 46(1):272–288.