Journal of Statistical Software

January 2014, Volume 56, Issue 6. *http://www.jstatsoft.org/*

MissMech: An R Package for Testing Homoscedasticity, Multivariate Normality, and Missing Completely at Random (MCAR)

Mortaza Jamshidian
California State
University, Fullerton

Siavash Jalal
University of California,
Los Angeles

Camden Jansen
University of California,
Irvine

Abstract

Researchers are often faced with analyzing data sets that are not complete. To properly analyze such data sets requires the knowledge of the missing data mechanism. If data are missing completely at random (MCAR), then many missing data analysis techniques lead to valid inference. Thus, tests of MCAR are desirable. The package **MissMech** implements two tests developed by Jamshidian and Jalal (2010) for this purpose. These tests can be run using a function called `TestMCARNormality`. One of the tests is valid if data are normally distributed, and another test does not require any distributional assumptions for the data. In addition to testing MCAR, in some special cases, the function `TestMCARNormality` is also able to test whether data have a multivariate normal distribution. As a bonus, the functions in **MissMech** can also be used for the following additional tasks: (i) test of homoscedasticity for several groups when data are completely observed, (ii) perform the k-sample test of Anderson-Darling to determine whether k groups of univariate data come from the same distribution, (iii) impute incomplete data sets using two methods, one where normality is assumed and one where no specific distributional assumptions are made, (iv) obtain normal-theory maximum likelihood estimates for mean and covariance matrix when data are incomplete, along with their standard errors, and finally (v) perform the Neyman's test of uniformity. All of these features are explained in the paper, including examples.

Keywords: Anderson-Darling, Hawkins, homogeneity of covariances, goodness of fit test, imputation, incomplete data, maximum likelihood estimate, missing data, Neyman's test, test of missing completely at random.

1. Introduction

In practice, often one is faced with analyzing data sets that are incomplete. It is well known that excluding cases that are incompletely observed from the analysis (complete case analysis) can lead to inefficient and/or biased inference. On the other hand, care must be taken to adopt methodologies that incorporate the incomplete cases, as the validity of such methods depends on the missing data mechanism, the process by which data have become incomplete. Rubin (1976) coined two popular terminologies of missing completely at random (MCAR) and missing at random (MAR) for two types of missing data mechanisms. In simple terms, MCAR is a process in which the missingness of the data is independent of both the observed and the missing values, and MAR is a process in which the missingness of the data depends on the observed values, but not on the missing values. When data are neither MCAR nor MAR, and in particular missingness depends on the missing data themselves, the missing data mechanism is called missing not at random (MNAR).

If data are MCAR, results from many missing data methods would be valid and complete case analysis of data will not lead to bias. On the other hand, if data are not MCAR, some missing data procedures may result in biased inference. Thus, to test for MCAR as a first step in the analysis of incomplete data is important. Little (1988) lists a number of important instances where verification of MCAR is important.

This paper reviews a few statistical tests of the MCAR missing data mechanism, and introduces the R (R Core Team 2013) package **MissMech** (Jamshidian, Jalal, and Jansen 2014) that implements state-of-the art MCAR tests developed by Jamshidian and Jalal (2010). As a by-product of the main routine, this package will be able to test for multivariate normality in some instances, and perform a number of other tests, as will be explained shortly.

Let $\mathbf{Y}$ be an n by p matrix of observations on n subjects and p variables with some of its elements missing. Furthermore, suppose that $\mathbf{Y}$ consists of g different missing (observed) data patterns, with the i-th missing data pattern consisting of n_i cases and p_i ($\leq p$) observed variables, for $i = 1, \ldots, g$; thus $n = \sum_{i=1}^{g} n_i$. Letting $\mathbf{\Sigma}_i$ denote the population covariance matrix for the i-th missing data pattern, in this paper we consider tests of MCAR that test the hypothesis

$$H_0 : \mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 = \ldots = \mathbf{\Sigma}_g \equiv \mathbf{\Sigma}, \tag{1}$$

namely, they test for homogeneity of covariances (homoscedasticity) between subsets (groups) of data having identical missing data patterns. The idea of dividing the data into identical missing data patterns for the purpose of testing MCAR goes back to Little (1988), where he proposed testing equality of means between the g missing data pattern groups, using a normal-theory likelihood ratio test. He argued that rejection of this test indicates that data are not MCAR. Little (1988) also mentioned a likelihood ratio test of homoscedasticity between the g groups, but noted that a large sample size would be required for this test to work well. In a simulation study, Kim and Bentler (2002) showed that the Little (1988) test of equality of means works well, however, their simulation confirmed Little's doubt that the likelihood ratio test for testing homoscedasticity fails in that the observed significance levels far exceed the nominal significance levels unless the sample size is quite large.

To test for MCAR, Kim and Bentler (2002) followed the same approach of testing equality of means and covariance between the g missing data patterns. Motivated by applications in structural equation models where covariances are modeled, they expressed importance of testing homoscedasticity, in addition to test of equality of means. They proposed tests based on least squares. Their tests had a few advantages over the likelihood ratio tests of Little (1988), namely they did not require that $n_i \geq p_i$ or that the data be normally distributed. In a simulation study, Kim and Bentler (2002) showed that their proposed methods perform better than the likelihood ratio tests of Little (1988), especially in testing homoscedasticity. Later Bentler, Kim, and Yuan (2004) noted that the Kim and Bentler (2002) test does not perform well when the n_is are small. This was confirmed by a simulation study performed by Jamshidian and Jalal (2010) who also shed some theoretical light on this shortcoming of the Kim and Bentler (2002) tests. Moreover, Jamshidian and Jalal (2010) showed that the Kim and Bentler (2002) test of homoscedasticity performs poorly when data are not normally distributed. They gave a theoretical argument, reasoning that if the population distribution is short tailed then the Kim and Bentler (2002) test of homoscedasticity's observed significance level would be below the nominal value, and if the population distribution is heavy-tailed then the observed significance level would be inflated as compared to the nominal value.

In the same vein as Little (1988) and Kim and Bentler (2002) and with the aim of improving on their tests, Jamshidian and Jalal (2010) proposed a normal-theory test and a nonparametric test of homoscedasticity to be used for testing for MCAR. The basis of the Jamshidian and Jalal (2010) tests is to impute missing data and employ complete data methods to test for homoscedasticity. More specifically, they adopt a test statistic proposed by Hawkins (1981) for testing homoscedasticity and normality of completely observed multivariate data. Based on various simulation studies, Jamshidian and Jalal (2010) showed that their proposed tests work better than those of Kim and Bentler (2002), both in terms of agreement of observed significance levels with their nominal counter part, and also in terms of power. This paper also elaborates on these points in Section 4.

As noted above, the **MissMech** implements the methods proposed by Jamshidian and Jalal (2010). This package's main aim is to test data for MCAR. However, as noted earlier, it can perform a number of other tasks including (i) test of multivariate normality in some instances, (ii) imputing missing data, (iii) testing for homoscedasticity and normality for complete data, (iv) performing k-sample test for equality of distribution between k groups of univariate data, (v) obtaining maximum likelihood estimates of the mean and covariance (including standard errors) for incomplete data, using the EM algorithm, and (vi) performing Neyman's smooth test of goodness of fit. Moreover, as we will see, multiple imputation is also available in the package as an exploratory tool to validate the result obtained by the MCAR test.

The remaining sections are organized as follows: In Section 2 we describe the two main tests of homoscedasticity used in the package **MissMech**, Section 3 gives a detailed account of the main function in **MissMech** called `TestMCARNormality` with several examples of the utility of this function, Section 4 discusses the performance of the tests in the package in terms of size and power, Section 5 describes a few important by-product functions in **MissMech**, and finally Section 6 gives a general discussion of the methodology and a few related R packages.

2. Tests of homoscedasticity, normality and MCAR

In this section we describe the statistical tests of Jamshidian and Jalal (2010) used to test MCAR and their implementation in the package **MissMech**. Let $\mathbf{Y}_i$ denote the n_i by p matrix of values corresponding to the i-th missing data pattern group in $\mathbf{Y}$, with $\mathbf{Y}_{\text{obs},i}$ and $\mathbf{Y}_{\text{mis},i}$ respectively denoting the observed and the missing part of $\mathbf{Y}_i$. Moreover, let

$\mathbf{Y}_{ij} = (\mathbf{Y}_{\text{obs},ij}, \mathbf{Y}_{\text{mis},ij})$ denote the j-th case in $\mathbf{Y}_i$, and $\mathbf{r}_{ij}$ denote a p by 1 vector of indicator variables with elements 1 and 0, respectively corresponding to the observed and missing values of $\mathbf{Y}_{ij}$. Assume that given $\mathbf{r}_{ij}$, $\mathbf{Y}_{ij}$ has the density $f(\mathbf{Y}_{ij}; \mathbf{\Sigma}_i, \boldsymbol{\theta})$ parameterized by the covariance matrix $\mathbf{\Sigma}_i = \mathsf{COV}(\mathbf{Y}_{ij})$ that depends on the missing data pattern i, and other parameters $\boldsymbol{\theta}$ that are assumed to be homogenous across missing data patterns. For example, $\boldsymbol{\theta}$ may include the mean or other types of parameters. Jamshidian and Jalal (2010) showed that under this setting, homogeneity of covariances implies MCAR. This is the premise underlying tests of MCAR that test homoscedasticity between various missing data patterns, as proposed by Little (1988), Kim and Bentler (2002), and Jamshidian and Jalal (2010).

The main goals in Jamshidian and Jalal (2010) were to test for MCAR, using tests of homoscedasticity, that worked well for data with small n_is as well as data that were not normally distributed. The former goal motivated them to utilize a test statistics that had been proposed by Hawkins (1981) and worked well for testing homoscedasticity in complete data when group sizes n_i were small. The Hawkins (1981) test assumes that a set of complete data $\mathbf{X}$ $(n \times p)$ from g groups is available, with $\mathbf{X}_{ij}$ denoting the j-th case from the i-th group; $j = 1, \ldots, n_i$ and $i = 1, \ldots, g$. He further assumes that $\mathbf{X}_{ij}$ follow a p-variate normal distribution with mean $\boldsymbol{\mu}_i$ and covariance $\mathbf{\Sigma}_i$. To test the hypothesis (1), he then proposes the use of the statistic F_{ij}, corresponding to case j in group i, defined by

$$F_{ij} = \frac{(n-g-p)n_i V_{ij}}{p\{(n_i-1)(n-g) - n_i V_{ij}\}}, \text{ where } V_{ij} = \left(\mathbf{X}_{ij} - \bar{\mathbf{X}}_i\right)^\top S^{-1}\left(\mathbf{X}_{ij} - \bar{\mathbf{X}}_i\right),$$

with $\bar{\mathbf{X}}_i$ and S, respectively denoting group i sample mean and the overall pooled sample covariance. Hawkins (1981) showed that, under (1) and the assumption of normality, the F_{ij}s follow a Snedecor's F distribution with degrees of freedom p and $n-g-p$. He proposed computing the statistic $A_{ij} = \mathsf{P}[\mathcal{F} > F_{ij}]$, the probability that an $\mathcal{F}$-distributed random variable with degrees of freedom p and $n-g-p$ exceeds F_{ij}. If the model of homoscedastic normal distribution holds, then A_{ij} is distributed as a uniform random variate over the range $(0,1)$. He proposed testing A_{ij} for uniformity as a test of homoscedasticity. Specifically, if A_{ij} are deemed not to be uniform on (0,1), then the null hypothesis (1) or the normality assumption are rejected.

Implementation of Hawkins' test requires complete data. For the problem at hand, where we are considering incomplete data, Jamshidian and Jalal (2010) proposed imputing the incomplete data and then applying the Hawkins' test to the completed data set. Two problems tacked by Jamshidian and Jalal (2010) for both normally and non-normally distributed data were first how to impute the data, and second how to utilize the F_{ij} statistics. In Sections 2.1 and 2.2 we will explain their solution to both of these problems and the methods that we implemented in the package **MissMech**.

2.1. Test of homoscedasticity under normality assumption

The normal-theory based test of homoscedasticity of Jamshidian and Jalal (2010) assumes that

$$\mathbf{Y}_{ij} = \begin{pmatrix} \mathbf{Y}_{\text{obs},ij} \\ \mathbf{Y}_{\text{mis},ij} \end{pmatrix} \sim \mathcal{N}_p(\boldsymbol{\mu}_i, \mathbf{\Sigma}_i) \equiv \mathcal{N}_p\left[\begin{pmatrix} \boldsymbol{\mu}_{o,i} \\ \boldsymbol{\mu}_{m,i} \end{pmatrix}, \begin{pmatrix} \mathbf{\Sigma}_{oo,i} & \mathbf{\Sigma}_{om,i} \\ \mathbf{\Sigma}_{mo,i} & \mathbf{\Sigma}_{mm,i} \end{pmatrix}\right].$$

Thus, from standard multivariate normal theory we have

$$\mathbf{Y}_{\text{mis},ij}|\mathbf{Y}_{\text{obs},ij}, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i \sim \mathcal{N}_{p-p_i}\left(\boldsymbol{\mu}_{m,i} + \boldsymbol{\Sigma}_{mo,i}\boldsymbol{\Sigma}_{oo,i}^{-1}(\mathbf{Y}_{\text{obs},ij} - \boldsymbol{\mu}_{o,i}), \boldsymbol{\Sigma}_{mm,i} - \boldsymbol{\Sigma}_{mo,i}\boldsymbol{\Sigma}_{oo,i}^{-1}\boldsymbol{\Sigma}_{om,i}\right). \tag{2}$$

As proposed by Jamshidian and Jalal (2010), random draws from this conditional distribution is used by **MissMech** to impute missing data. To generate the random draws, values for $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$ are needed. Since the aim is to test the hypothesis (1), Jamshidian and Jalal (2010) assumed $\boldsymbol{\mu}_1 = \ldots = \boldsymbol{\mu}_g \equiv \boldsymbol{\mu}$ and $\boldsymbol{\Sigma}_1 = \ldots = \boldsymbol{\Sigma}_g \equiv \boldsymbol{\Sigma}$ and proposed to estimate the common mean $\boldsymbol{\mu}$ and the common covariance $\boldsymbol{\Sigma}$ using the method of maximum likelihood (see, e.g., Jamshidian and Bentler 1999). This method is implemented in **MissMech**. If the means are not equal, then ML estimates of $\boldsymbol{\mu}_i$ can be used for each group. The option `ImputationMethod = "Normal"` in the function `TestMCARNormality` within the package **MissMech** uses the above method to impute missing data. This is not the default method of imputation in the package, however.

Once the missing data are imputed, the statistics A_{ij} need to be computed and tested for uniformity. In the package **MissMech**, we follow Jamshidian and Jalal (2010)'s recommendation of obtaining, for each of the groups $i = 1, \ldots, g$, a p value P_i that is used to test uniformity of A_{ij} for group i. We then use the Fisher (1932) result

$$P_T = \sum_{i=1}^{g}(-2\log P_i) \sim \chi^2_{2g} \tag{3}$$

to combine the P_is to obtain an overall p value for testing uniformity of all A_{ij}.

To compute the P_is, Jamshidian and Jalal (2010) proposed using the Neyman (1937) test for uniformity. Their recommendation was based on a simulation study that they performed to compare a few different tests of uniformity, including the Anderson and Darling (1954) test that was used by Hawkins (1981). The Neyman's test statistic is

$$N_{ik} = \sum_{\ell=1}^{k}\left\{n_i^{-1/2}\sum_{j=1}^{n_i}\pi_\ell(A_{ij})\right\}^2 \quad i = 1, \ldots, g, \tag{4}$$

where $\pi_1, \pi_2, \ldots, \pi_k$ are normalized Legendre polynomials on $(0,1)$. As noted by Jamshidian and Jalal (2010), a value of $k = 4$ works well in practice and this is the value that we use in **MissMech**. Large values of the test statistic in (4) point to rejection of uniformity. To obtain a P_i, the **MissMech** package simulates a large number of N_{ik}s by simulating a large number of $A_{ij} \sim \text{Uniform}(0,1)$, and computes the proportion of simulated N_{ik}s that are larger than the N_{ik} obtained from the data. In cases where n_is are large, the statistics N_{ik} have an approximate χ^2 distribution with k degrees of freedom. For these cases, the **MissMech** package allows the user to set a threshold value n_{min} such that if $n_i \geq n_{min}$, then the χ^2 approximation will be used in place of the simulated distribution for the i-th group. The main advantage of using the χ^2 distribution is that it is much less computationally demanding than the simulation method. On the other hand, the p values obtained based on the simulation method are usually more accurate, especially when the n_is are small and the number of simulated N_{ik} is large. We compared the χ^2 approximation to the simulation method, using a simulation study. Based on this study, we recommend using $n_{min} = 30$ for the asymptotic method and using 10,000 simulated values for the simulation-based method. These are the default values in the **MissMech** package.

2.2. The nonparametric test

As noted above, the nonparametric test of homoscedasticity of Jamshidian and Jalal (2010) assumes that data come from a density of the form $f(\mathbf{Y}_{ij}; \mathbf{\Sigma}_i, \boldsymbol{\theta})$ and tests equality of the covariances $\mathbf{\Sigma}_i$. To perform this test, Jamshidian and Jalal (2010)'s nonparametric test computes the Hawkins test statistic F_{ij} as described above. Obviously, if the data distribution is unknown, the distribution of F_{ij} is also not known. Jamshidian and Jalal (2010), however, showed that if the data have a density of the form $f(\mathbf{Y}_{ij}; \mathbf{\Sigma}_i, \boldsymbol{\theta})$ and the n_is are equal or large, then under the assumption of homoscedasticity (1) the distribution of F_{ij} for all the g groups must be identical. They propose to take advantage of this fact to test for homoscedasticity. Two things need close attention here: first, and as before, computation of F_{ij} requires imputation of the data and this imputation must be done without making any distributional assumptions about the data. Second, an appropriate k-sample test must be employed to test equality of distribution of F_{ij} in the g groups.

As proposed by Jamshidian and Jalal (2010), the package **MissMech** employs a method of imputation, in the spirit of that proposed by Srivastava and Dolatabadi (2009), that only assumes independence of observations from case to case and the continuity of their cumulative distribution function. Details of the imputation method is given by Jamshidian and Jalal (2010, Section 3.2). In short, imputation values are obtained by adding an appropriate random error to the best linear predictors of the missing observations. To employ this method, an estimate of the mean and covariance for the variables is required. For this case, these estimates are obtained from the completely observed cases. In the package **MissMech** the option `ImputationMethod = "Dist.Free"` triggers this method of imputation, and in fact this option is the default option for the package. It should be noted that, to obtain a reasonable estimate of the mean and covariance, one has to have a sufficient number of completely observed cases. As such, if the number of completely observed cases is less than $\min(10, 2p)$, then **MissMech** gives a warning and reverts to `ImputationMethod = "Normal"`. The value $\min(10, 2p)$ was set based on our experience, and can be changed within the code, if desired.

Jamshidian and Jalal (2010) discuss various k-sample tests to test equality of distribution of F_{ij} for the g groups. In the package **MissMech**, we follow their recommendation and use the Scholz and Stephens (1987) test, also known as the Anderson-Darling k-sample test. This test uses a rank statistic of the form $T = \frac{1}{n}\sum_{i=1}^{g} T_i$ with

$$T_i = \frac{1}{n_i}\sum_{j=1}^{n-1}\frac{(nM_{ij} - jn_i)^2}{j(n-j)}, \tag{5}$$

where M_{ij} is the number of observations in the i-th sample that are not greater than the j-th order statistic in the pooled sample of F_{ij}s.

In summary, the nonparametric test imputes the missing data using the Srivastava and Dolatabadi (2009) approach, computes F_{ij} as defined above, and applies the Anderson-Darling k-sample test for equality of distribution of F_{ij}s amongst groups $i = 1, \ldots, g$. If this test is rejected, it will be concluded that the covariances are non-homogeneous, and data are not MCAR.

3. The MissMech package

The **MissMech** package consists of several functions that can be independently employed, with the function `TestMCARNormality` being the main function. A description of each of the functions in the package can be obtained by the R function `help`. In the subsections that follow, however, we provide a more detailed description of the functions, including examples and reference to the technical notes made above.

3.1. The function `TestMCARNormality`

The syntax, including various options, for the function `TestMCARNormality` is as follows:

```
TestMCARNormality(data, del.lesscases = 6, imputation.number = 1,
  method = "Auto", imputation.method = "Dist.Free", nrep = 10000,
  n.min = 30, seed = 110, alpha = 0.05, imputed.data = NA)
```

Below, we provide a detailed description of this function.

`data` is an input to this function consisting of a data set with incompletely observed cases. This data set must be a matrix or data frame with missing data coded as `NA`.

`del.lesscases` is an option that allows the user to remove, from the analysis, missing data patterns with n_i less than or equal to the number of cases specified in this option. As explained in Jamshidian and Jalal (2010, Section 5), the larger the n_is the better their tests perform. However, their tests perform well for n_i as small as 3, and their simulation shows great performance when $n_i > 6$. So, we have set the default `del.lesscases = 6`, and the user has the option of changing this value. Based on our experience, a value as small as `del.lesscases = 3` works well. The smallest value allowed, however, is 1, since for our analysis we require at least two cases in each of the missing data patterns. Indeed if there are no groups with small n_is, this option would not be useful.

`imputation.number` specifies the number of times the missing data are imputed. This option allows the user to impute the data multiple times. The default for this option is 1, as the test results are based on a single imputation. However, since the imputation procedures include random components, the statistical test results can vary depending on the random values used in the imputation process. The multiple imputation option allows the user to examine the performance of the test under different imputation values. More specifically, for each imputation, this option produces the P_i values for $i = 1, \ldots, g$, based on the statistic (3), and in the case that the nonparametric test is used, T_i values, given in (5), are also reported. Note that, small values of P_i or large values of T_i support rejection of H_0. The package **MissMech** can generate a boxplot of the P_i and T_i, using the output of the `TestMCARNormality` function. We will give more details on utility of such boxplots in Example 4 in Section 3.2.

`method` is an option that allows the user to select either the Hawkins or the nonparametric method for the test. If the user believes that the data follow multivariate normal distribution, `method = "Hawkins"` should be selected. On the other hand, if data are not normally distributed, then `method = "Nonparametric"` should be used. If the user is unsure, then the default value of `method = "Auto"` will be used, in which case both the Hawkins and the nonparametric tests will be run, and the default output follows the recommendation by Jamshidian and Jalal (2010) outlined in their flowchart given in Figure 7 of their paper.

`imputation.method` is an option that allows the user to choose one of the two imputation methods described in Sections 2.1 and 2.2. The option `"Normal"` will generate imputed values based on the conditional distribution (2), and the option `"Dist.Free"` will generate imputation values based on the method of Srivastava and Dolatabadi (2009) described in Section 2.2. The latter is the default imputation method and, as noted above, requires $\min(10, 2p)$ number of complete cases. If this number of complete cases is not available, the program gives a warning and uses the method `"Normal"` to impute the data. We have selected the value of $\min(10, 2p)$ based on our experience, and with the rationale that there must be sufficient number of complete cases to estimate the population covariance matrix, as required by the the method of Srivastava and Dolatabadi (2009).

`nrep` is an option that allows the user to set the number of replications used to simulate the empirical distribution of the statistic N_{ik} in (4) under the null. This empirical distribution is used to obtain critical values for the Neyman test of uniformity. Clearly the larger the number of replications, the more accurate the critical value would be. As noted in Section 2.2, the default value is set to be 10,000. This leads to a reasonable accuracy in a reasonable time. If the evidence, however, is marginal in terms of the p value obtained, one may increase this default value to increase accuracy. See also the option `n.min`.

`n.min` is an option that allows the user to set a value for the $n_{\min}$, described in Section 2.1. Specifically, if for a given group the sample size n_i is at least as large or larger than the value set by `n.min`, then the asymptotic χ^2 distribution will be used for the N_{ik} in the Neyman test of uniformity for that group. Clearly, if `n.min` is set to $\max_i n_i + 1$, then the simulation method will be used for all groups, and if `n.min` is set to $\min_i n_i$, the asymptotic approximation will be used for all groups. Also note that in the cases where `imputation.number` is set to a value larger than 1, and the simulation method is used to obtain an empirical distribution of N_{ik}, it suffices to simulate the null values of N_{ik} only once, and this is the case in the **MissMech** package.

`seed` is an option that allows the user to set a seed for random number generation. The default seed is 110. If the value is set to `NA`, a system selected seed is used. Note that random numbers are used to obtain imputation values as to obtain an empirical distribution for N_{ik}.

`alpha` is an option that allows the users to set the significance level at which the statistical tests are performed. The default value is 0.05.

`imputed.data` is an option that allows the user to input an imputed data set of their own. When this option is used, `data` must be a matrix or data frame with incomplete data prior to imputation and `imputed.data` must be the corresponding data set (with the same order of cases) with incomplete data filled by a method of user's choice. When `imputed.data` is specified, neither the imputation methods of `"Normal"` nor `"Dist.Free"` will be used.

3.2. Examples

In this section we give a few examples of applications of the **MissMech** package.

Example 1: MCAR and normal data

We start with an example where we generate $n = 300$ cases from the multivariate normal $\mathcal{N}_5(0, I)$. We then remove each datum, independently of the other data, with probability of 20%. The result is an incomplete data set with data that are MCAR. We then apply the `TestMCARNormality` function.

```
R> n <- 300
R> p <- 5
R> pctmiss <- 0.2
R> set.seed(1010)
R> y <- matrix(rnorm(n * p), nrow = n)
R> missing <- matrix(runif(n * p), nrow = n) < pctmiss
R> y[missing] <- NA
R> out <- TestMCARNormality(data = y)
R> out

Call:
TestMCARNormality(data = y)

Number of Patterns:  9

Total number of cases used in the analysis:  245

 Pattern(s) used:
                                   Number of cases
group.1    1    1    1   NA    1                16
group.2    1    1   NA    1    1                23
group.3    1    1    1    1    1               101
group.4   NA    1    1    1    1                33
group.5    1   NA    1    1    1                22
group.6   NA   NA    1    1    1                 7
group.7    1    1    1    1   NA                21
group.8    1    1   NA    1   NA                10
group.9   NA    1    1    1   NA                12


    Test of normality and Homoscedasticity:
  -------------------------------------------

Hawkins Test:

    P-value for the Hawkins test of normality and homoscedasticity: 0.948
    There is not sufficient evidence to reject normality
    or MCAR at 0.05 significance level
```

The output includes the number of imputations (which by default is 1), the number of missing data patterns (in this example 9), and the number of cases in each of the patterns. For example, `group.1` consists of 16 cases that have missing values on the fourth variable only. Also, `group.3` includes 101 cases that are completely observed. Furthermore, a total of $16 + 23 + \ldots + 12 = 245$ out of the 300 cases generated is used in the analysis. The 55 cases that were excluded belonged to groups of missing data patterns with 6 or less cases in each group. For this example, the p value for the Hawkins test is 0.948, from which we infer that there is not sufficient evidence to reject either the normality or the MCAR assumptions. This is expected for this example, as data were generated from a normal distribution and the missing data mechanism was MCAR.

When the p value for the Hawkins test is larger than the value of `alpha` specified, neither the hypothesis of normality nor the hypothesis of MCAR is rejected, and thus the program does not report the result of the nonparametric test. However, users may obtain the p value for the nonparametric test by applying the `summary` function, as shown below. For this example, the p value for the nonparametric test is 0.711.

```
R> Out <- TestMCARNormality(data = y)
R> summary(Out)
```

We can include more cases, from the 300 cases, in the analysis of the above data, by setting `del.lesscases = 1`, namely using

```
R> out1 <- TestMCARNormality(data = y, del.lesscases = 1)
R> out1

Call:
TestMCARNormality(data = y, del.lesscases = 1)

Number of Patterns:  22

Total number of cases used in the analysis:  293

 Pattern(s) used:
                                                Number of cases
group.1      1    1    1   NA    1                           16
group.2      1    1   NA    1    1                           23
group.3      1    1    1    1    1                          101
group.4     NA    1    1    1    1                           33
group.5      1   NA    1    1    1                           22
group.6     NA   NA    1    1    1                            7
group.7      1    1    1   NA   NA                            5
group.8      1    1    1    1   NA                           21
group.9      1   NA    1   NA    1                            5
group.10     1    1   NA    1   NA                           10
group.11    1   NA    1    1   NA                  5
group.12   NA    1    1    1   NA                 12
group.13    1   NA   NA   NA    1                  4
group.14   NA    1    1   NA    1                  5
group.15   NA   NA    1    1   NA                  3
group.16    1   NA   NA    1    1                  5
group.17    1    1   NA   NA   NA                  2
group.18   NA    1   NA   NA    1                  3
group.19   NA   NA    1   NA    1                  2
group.20   NA    1    1   NA   NA                  2
group.21   NA    1   NA    1    1                  5
group.22    1    1   NA   NA    1                  2


    Test of normality and Homoscedasticity:
  -------------------------------------------

Hawkins Test:

    P-value for the Hawkins test of normality and homoscedasticity: 0.279

    There is not sufficient evidence to reject normality
    or MCAR at 0.05 significance level
```

The output indicates that the number of data patterns increases to 22, and the number of cases used in the analysis increases to a total of 293. The remaining 7 cases were in missing data patterns that had a single case. Again, as expected, the hypotheses of normality and homoscedasticity are not rejected based on the p value of 0.279.

To see how randomness of the imputations affects the results, we apply the multiple imputation option and look at the boxplot of the p values. This is done by issuing the following commands:

```
R> Out <- TestMCARNormality(data = y, imputation.number = 100)
R> summary(Out)
R> boxplot(Out)
```

The top panel of Figure 1 shows boxplots of 100 P_i values for each of the 9 groups. The dashed red line indicates the cut-off value, which is set to α/g, in this case it is set to `alpha = 0.05/9`. While the p values vary, they all are above the shown cut-off line. Note that group 3 is the group of completely observed data with the largest number of cases, namely 101, and smallest variation in the p values. The bottom panel of Figure 1 shows the boxplots corresponding to 100 T_i values for the nonparametric test for each of the groups. While the T_i scales are not easily interpretable, this plot can serve as an exploratory tool to identify possible outlying groups (i.e., groups with exceptionally large T_is), which in this example do not exist.

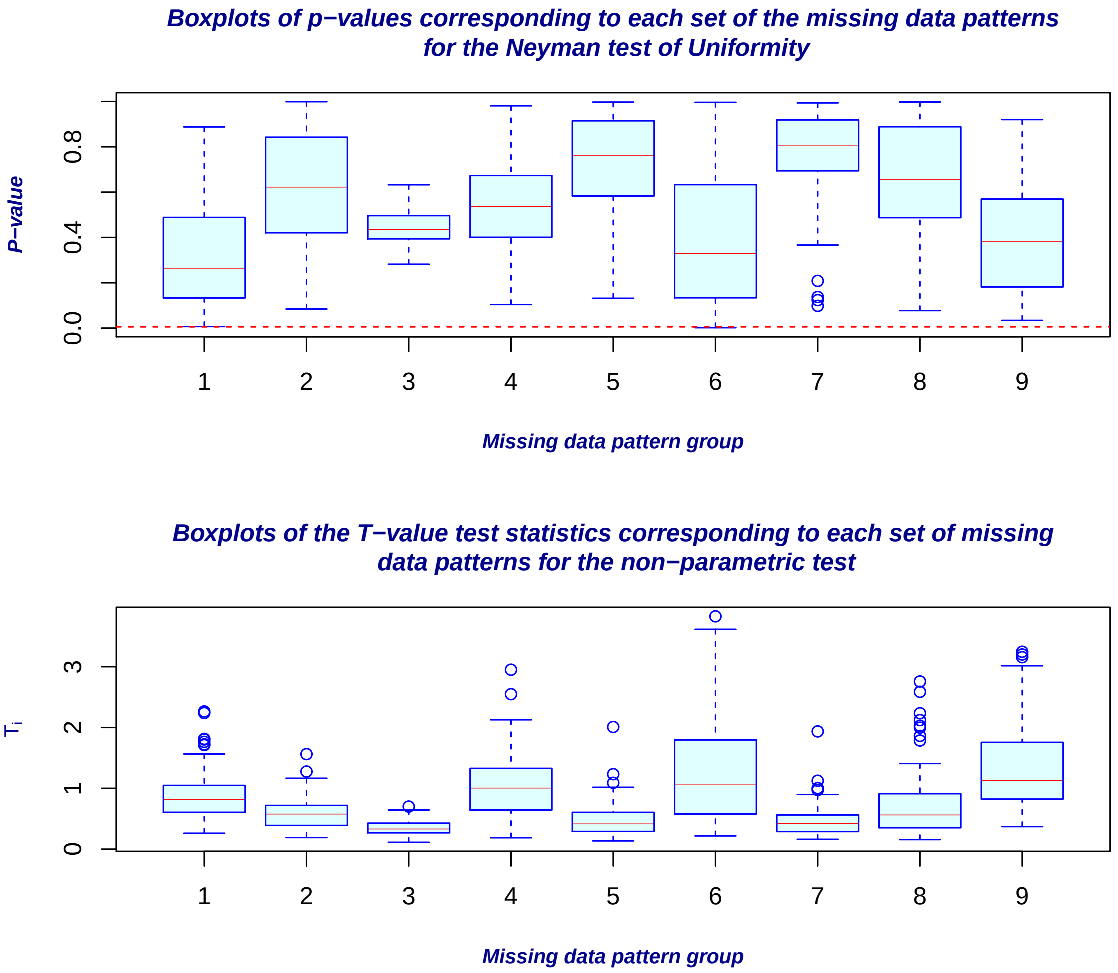


Figure 1: Result of 100 multiple imputations in Example 1.

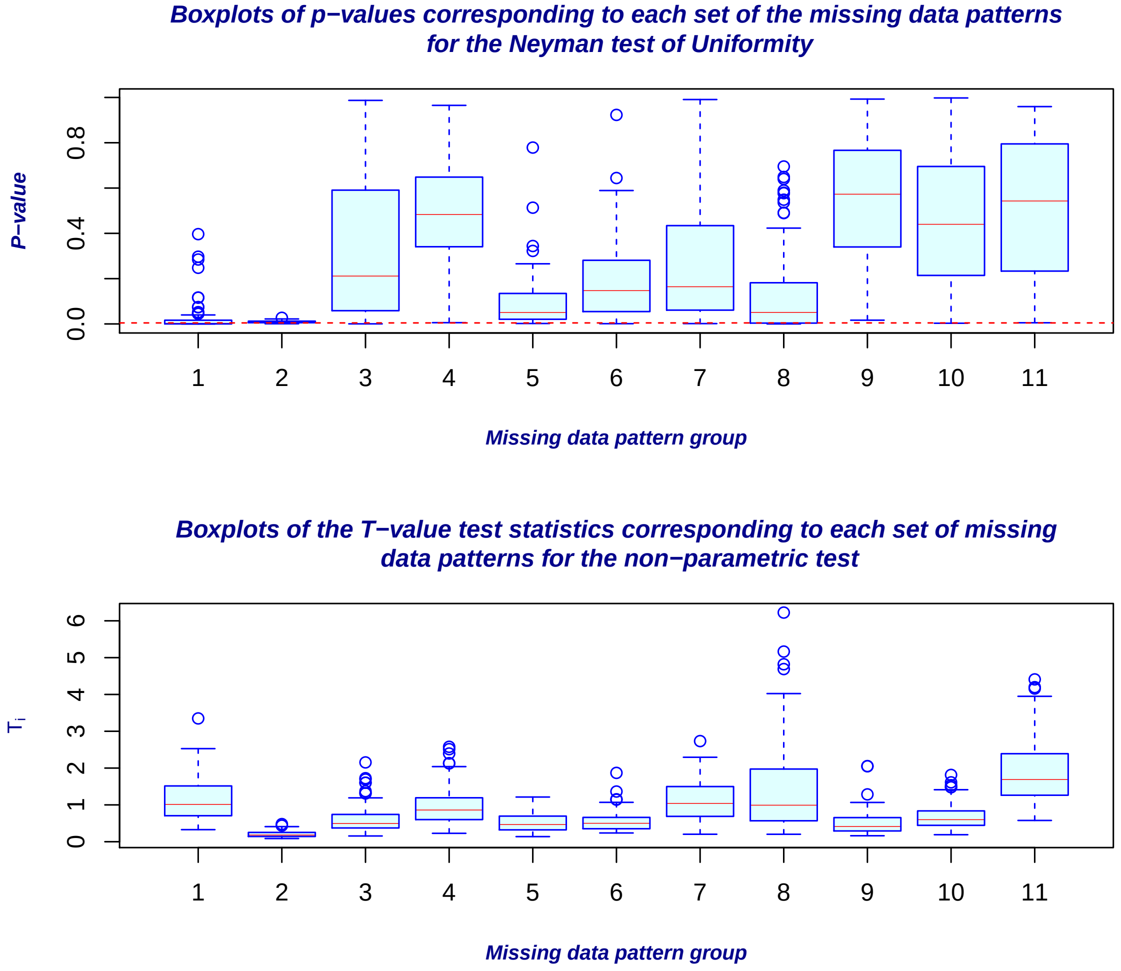
Example 2: MCAR and non-normal data

This example uses data that are not normally distributed, but are MCAR. Specifically, we generate independent data from the t distribution with 5 degrees of freedom, and apply the `TestMCARNormality` function.

```
R> n <- 300
R> p <- 5
R> pctmiss <- 0.2
R> set.seed(1010)
R> y <- matrix(rt(n * p, 5), nrow = n)
R> missing <- matrix(runif(n * p), nrow = n) < pctmiss
R> y[missing] <- NA
R> out <- TestMCARNormality(data = y)
R> out
```

```
Call:
TestMCARNormality(data = y)

Number of Patterns:  11

Total number of cases used in the analysis:  258

 Pattern(s) used:
                                    Number of cases
group.1    NA    1    1    1    1                31
group.2     1    1    1    1    1               100
group.3     1    1    1   NA   NA                12
group.4     1    1    1   NA    1                20
group.5     1    1    1    1   NA                19
group.6     1   NA    1    1    1                23
group.7    NA   NA    1    1    1                 7
group.8     1   NA    1   NA    1                10
group.9     1    1   NA    1    1                20
group.10    1    1   NA    1   NA                 8
group.11   NA    1    1   NA    1                 8


    Test of normality and Homoscedasticity:
  -------------------------------------------

Hawkins Test:

    P-value for the Hawkins test of normality and homoscedasticity: 0.000266

    Either the test of multivariate normality or homoscedasticity (or both)
    is rejected.
    Provided that normality can be assumed, the hypothesis of MCAR is
    rejected at 0.05 significance level.

Non-Parametric Test:

    P-value for the non-parametric test of homoscedasticity: 0.51

    Reject Normality at 0.05 significance level.
    There is not sufficient evidence to reject MCAR at 0.05 significance
    level.
```

For this example, the p value for the Hawkins test is small (0.0003) and this is evidence that either normality or homoscedasticity (or both) is rejected. This test is then followed by the nonparametric test of homoscedasticity which results in a high p value (0.51) from which we conclude that there is no evidence of heteroscedasticity. Thus, as it should ideally happen, the software concludes that there is evidence of non-normality, but no evidence that data are


Boxplots of p−values corresponding to each set of the missing data patterns for the Neyman test of Uniformity

Boxplots of the T−value test statistics corresponding to each set of missing data patterns for the non−parametric test

Figure 2: Result of 100 multiple imputations for Example 2.

not MCAR.

Figure 2 shows the boxplots corresponding to a multiple imputation run, based on 100 imputations, for Example 2. Interestingly, with exception of a few outliers, the p values for groups 1 and 2 are small. These groups are the two groups with the largest number of cases. The T_i value for the nonparametric test is smallest for group 2, the largest group, but it does not seem to distinguish itself significantly from the other groups.

As noted above, one may impute the data using a method other than the methods available in the package **MissMech**. In the following, we use the k nearest neighbor method of the package **imputation** (Wong 2013) to impute the missing data of Example 2 and then run the **MissMech** tests.

```
R> library("imputation")
R> yimputed <- kNNImpute(y, k = 3)
R> out <- TestMCARNormality(data = y, imputed.data = yimputed$x)
R> out
```

```
Call:
TestMCARNormality(data = y, imputed.data = yimputed$x)

Number of Patterns:  11

Total number of cases used in the analysis:  258

 Pattern(s) used:
                                    Number of cases
group.1    NA    1    1    1    1                31
group.2     1    1    1    1    1               100
group.3     1    1    1   NA   NA                12
group.4     1    1    1   NA    1                20
group.5     1    1    1    1   NA                19
group.6     1   NA    1    1    1                23
group.7    NA   NA    1    1    1                 7
group.8     1   NA    1   NA    1                10
group.9     1    1   NA    1    1                20
group.10    1    1   NA    1   NA                 8
group.11   NA    1    1   NA    1                 8


    Test of normality and Homoscedasticity:
  -------------------------------------------

Hawkins Test:

    P-value for the Hawkins test of normality and homoscedasticity: 1.17e-13

    Either the test of multivariate normality or homoscedasticity (or both)
    is rejected.
    Provided that normality can be assumed, the hypothesis of MCAR is
    rejected at 0.05 significance level.

Non-Parametric Test:

    P-value for the non-parametric test of homoscedasticity: 0.00408

    Hypothesis of MCAR is rejected at  0.05 significance level.
    The multivariate normality test is inconclusive.
```

Example 3: Normally distributed data and not MCAR

In this example, we generate data from a multivariate normal distribution with mean zero, variance one, and all correlations equal to 0.3. The missing data mechanism imposed is MAR (not MCAR) and it was implemented by setting each $\mathbf{Y}_{ij}$ for $i = 1, \ldots, n$ and $j = 2, \ldots, p$ to missing provided that $\mathbf{Y}_{i,j-1} > 0.8$. The value 0.8 was selected to achieve approximately 15 to 20 percent missing values.

```
R> n <- 300
R> p <- 5
R> r <- 0.3
R> mu <- rep(0, p)
R> sigma <- r * (matrix(1, p, p) - diag(1, p)) + diag(1, p)
R> set.seed(110)
R> eig <- eigen(sigma)
R> sig.sqrt <- eig$vectors %*% diag(sqrt(eig$values)) %*% solve(eig$vectors)
R> sig.sqrt <- (sig.sqrt + t(sig.sqrt)) / 2
R> y <- matrix(rnorm(n * p), nrow = n) %*% sig.sqrt
R> tmp <- y
R> for (j in 2:p)
+    y[tmp[, j - 1] R > 0.8, j] <- NA
R> out <- TestMCARNormality(data = y, alpha = 0.1)
R> out
```

```
Call:
TestMCARNormality(data = y, alpha = 0.1)

Number of Patterns:  9

Total number of cases used in the analysis:  277
 Pattern(s) used:
                                  Number of cases
group.1   1    1    1   NA    1                19
group.2   1   NA    1    1    1                21
group.3   1    1    1    1    1               163
group.4   1    1   NA   NA    1                 8
group.5   1   NA    1    1   NA                 7
group.6   1   NA    1   NA   NA                 7
group.7   1    1   NA    1    1                25
group.8   1    1    1    1   NA                20
group.9   1    1   NA    1   NA                 7


    Test of normality and Homoscedasticity:
  -------------------------------------------

Hawkins Test:

    P-value for the Hawkins test of normality and homoscedasticity: 0.053

    Either the test of multivariate normality or homoscedasticity (or both)
    is rejected.
    Provided that normality can be assumed, the hypothesis of MCAR is
    rejected at 0.1 significance level.

Non-Parametric Test:

    P-value for the non-parametric test of homoscedasticity: 0.000341

    Hypothesis of MCAR is rejected at  0.1 significance level.
    The multivariate normality test is inconclusive.
```

For this example, the Hawkins test is rejected at the 10% significance level (p value = 0.053), pointing to either non-normality or heteroscedasticity. Note that we set the option `alpha = 0.1` to force the program to output the result of the nonparametric test. The nonparametric test rejects the null hypothesis of homoscedasticity. Thus, it is concluded that there is sufficient evidence that data are not MCAR. Note that rejection of Hawkins test, in the absence of any distributional assumptions about the data, simply points to either non-normality or heteroscedasticity which is indistinguishable by the test. Thus, in this case we can make no conclusion about non-normality of the data.

We examined the power of the Hawkins and the nonparametric tests by simulating 1000 sets of data, using the same scheme of data generation used in this example. In 1000 runs, the Hawkins test rejected H_0 approximately 54.7% of the time and the nonparametric test rejected the null hypothesis, approximately 83.9% of the time, indicating especially good power for the nonparametric test. In the next section we discuss results of a few more simulation studies to shed some light on the power of the Hawkins and nonparametric tests.

Example 4: Identifying group(s) that differ

In this example we show how multiple imputation and boxplots can be used to identify one or two groups that have a different covariance matrix than others. To show this, we generated data from standard normal and created missingness based on the MCAR missing data mechanism, as in Example 1. Then, we identified the two largest groups, not including the complete data group, and multiplied the generated values for these groups by 2, thus making the covariance matrices for these two groups different from the other groups. The code involves use of the **MissMech** function `OrderMissing`, which orders the missing data according to their missing data patterns.

```
R> n <- 300
R> p <- 5
R> pctmiss <- 0.2
R> set.seed(1010)
R> y <- matrix (rnorm(n * p), nrow = n)
R> missing <- matrix(runif(n * p), nrow = n) < pctmiss
R> y[missing] <- NA
R> Out <- OrderMissing(y)
R> y <- Out$data
R> spatcnt <- Out$spatcnt
R> g2 <- seq(spatcnt[1] + 1, spatcnt[2])
```

```
R> g4 <- seq(spatcnt[3] + 1, spatcnt[4])
R> y[c(g2, g4), ] <- 2 * y[c(g2, g4), ]
R> out <- TestMCARNormality(data = y, imputation.number = 100)
R> out
R> boxplot(out)

Call:
TestMCARNormality(data = y, imputation.number = 100)

Number of Patterns:  9

Total number of cases used in the analysis:  245

 Pattern(s) used:
                                   Number of cases
group.1    1    1    1   NA    1                16
group.2    1    1   NA    1    1                23
group.3    1    1    1    1    1               101
group.4   NA    1    1    1    1                33
group.5    1   NA    1    1    1                22
group.6   NA   NA    1    1    1                 7
group.7    1    1    1    1   NA                21
group.8    1    1   NA    1   NA                10
group.9   NA    1    1    1   NA                12

    Test of normality and Homoscedasticity:
  -------------------------------------------

Hawkins Test:

    P-value for the Hawkins test of normality and homoscedasticity: 3.09e-15

    Either the test of multivariate normality or homoscedasticity (or both)
    is rejected.
    Provided that normality can be assumed, the hypothesis of MCAR is
    rejected at 0.05 significance level.

Non-Parametric Test:

    P-value for the non-parametric test of homoscedasticity: 1.01e-33

    Hypothesis of MCAR is rejected at  0.05 significance level.
    The multivariate normality test is inconclusive.
```

The groups that have covariance $4I$ are `group.2` and `group.4`. As expected, both the Hawkins test and the MCAR test are rejected due to non-homogeneous covariances. Inspecting the boxplot obtained by multiple imputation, shown in Figure 3, we see that the boxplots for

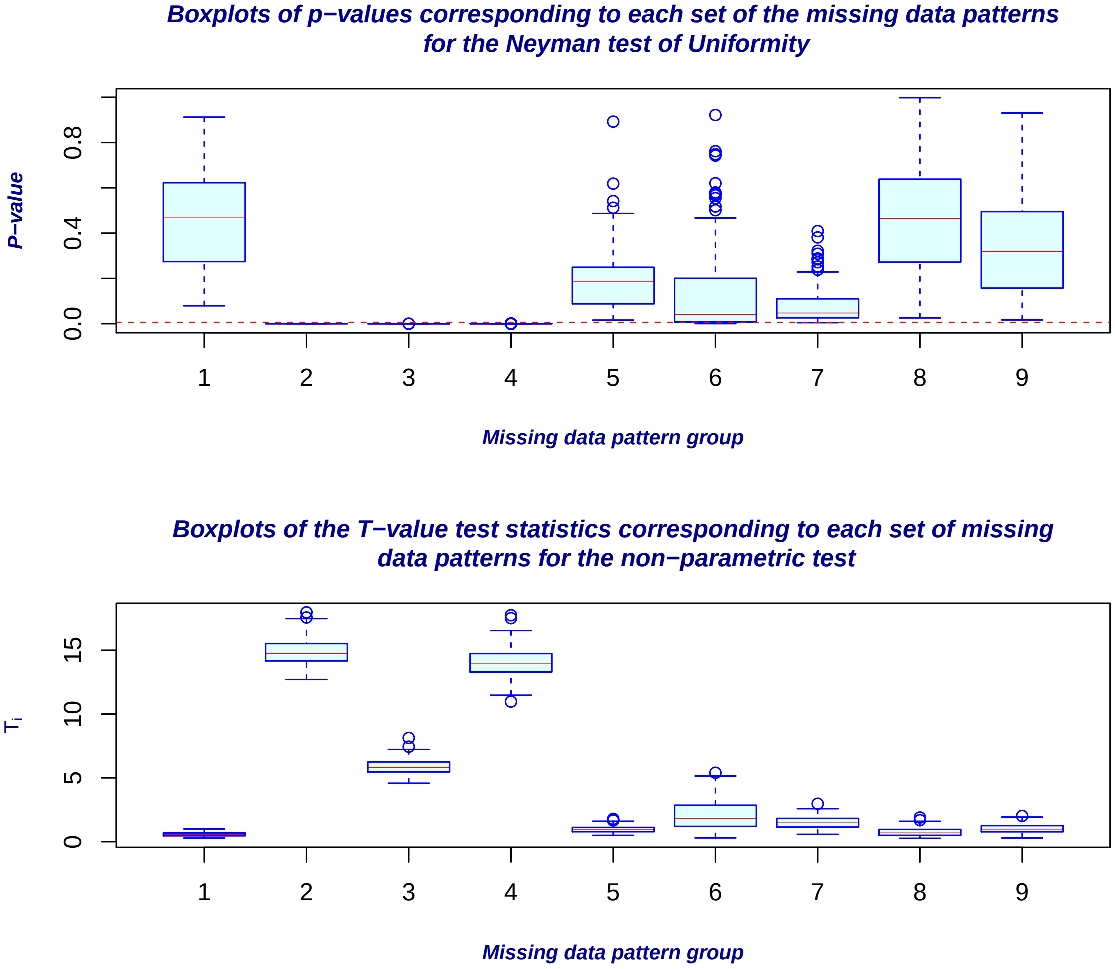


Figure 3: Result of 100 multiple imputations for Example 4.

group.2, group.3, and group.4 stand out. In particular, the boxplot of T_i values for group.2 and group.4 are highest, thus identifying these two special groups. Note that group.3 is the group with complete cases and has a much larger number of cases than the other groups.

In a next step to analyze these data, we use the code shown below to remove group 2, and apply the analysis to the newly formed data set.

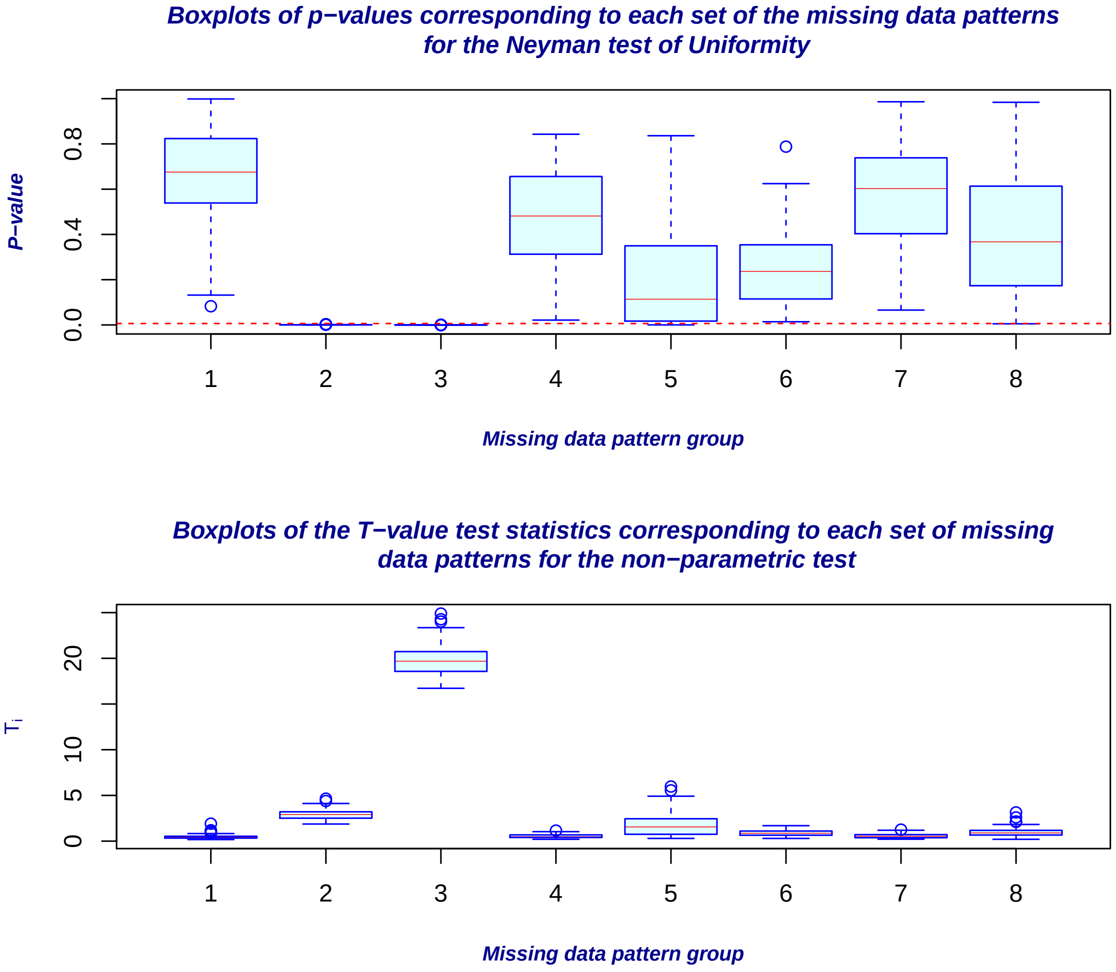
```
R> y1 <- y[-seq(spatcnt[1] + 1, spatcnt[2]), ]
R> Out <- TestMCARNormality(data = y1, imputation.number = 100)
R> boxplot(Out)

Call:
TestMCARNormality(data = y1, imputation.number = 100)

Number of Patterns:  8
```

```
Total number of cases used in the analysis:  222

 Pattern(s) used:
                                   Number of cases
group.1    1    1    1   NA    1                16
group.2    1    1    1    1    1               101
group.3   NA    1    1    1    1                33
group.4    1   NA    1    1    1                22
group.5   NA   NA    1    1    1                 7
group.6    1    1    1    1   NA                21
group.7    1    1   NA    1   NA                10
group.8   NA    1    1    1   NA                12


    Test of normality and Homoscedasticity:
  -------------------------------------------

Hawkins Test:

    P-value for the Hawkins test of normality and homoscedasticity: 5.38e-17

    Either the test of multivariate normality or homoscedasticity (or both)
    is rejected.
    Provided that normality can be assumed, the hypothesis of MCAR is
    rejected at 0.05 significance level.

Non-Parametric Test:

    P-value for the non-parametric test of homoscedasticity: 4.72e-15

    Hypothesis of MCAR is rejected at  0.05 significance level.
    The multivariate normality test is inconclusive.
```

Now, the covariance for `group.3` (formerly `group.4`), is different from all other groups. The boxplot corresponding to this run is given in Figure 4, with the T_i values for `group.3` clearly standing out, thus identifying the group that has a different covariance matrix. We should note that this type of analysis simply has exploratory value, and once a group is identified as possibly having a different covariance matrix further confirmatory analysis need to be performed to make definitive conclusions.

Example 5: Test of homoscedasticity for complete data

While the main use of the `TestMCARNormality` function is to test MCAR for an incomplete data set, this function can also be utilized for test of homoscedasticity between several groups based on completely observed data. A way to accomplish this is to set the option `imputed.data` equal to the name of the complete data set for which homoscedasticity is to be tested, and identify groups via artificial missing data patterns. More specifically, we as-


Boxplots of p−values corresponding to each set of the missing data patterns for the Neyman test of Uniformity

Boxplots of the T−value test statistics corresponding to each set of missing data patterns for the non−parametric test

Figure 4: Result of 100 multiple imputations in Example 4 after removing group 2.

sume a unique missing data pattern for each group (e.g., `1 NA 1 1 1`), and build a data set corresponding to the complete data in which for each case we include the designated missing data pattern for the group that the case belongs to. We input this constructed data matrix for `data` in `TestMCARNormality`.

In this example we generate three groups of data from the multivariate normal distribution with mean zero and covariances as follows: Group 1 has all variances equal to 1 and covariances equal to 0.2, and groups 2 and 3 have I and $2I$ as their covariance matrices, respectively. The complete data set called `ycomplete` is assigned to the option `imputed.data`. Thus the program will use this data set to test for homoscedasticity. The data set `ygroup` is a data set which has the same size as the data set `ycomplete`, with cases in group 1 replaced by `NA 1 1 1 1`, cases in group 2 replaced by `NA 1 NA 1 1`, and finally cases in group 3 replaced by `1 NA 1 1 1`. The pattern designation using `1`s and `NA`s can be done in any arbitrary way, as long as each group is assigned a unique pattern. Note that this allows designation of $2^p - 1$ groups, if we have p variables in the data set. For this problem, we have used the option `method = "Hawkins"` since we know that the data are normally distributed.

```
R> n <- 50
R> p <- 5
R> r <- 0.4
R> sigma <- r * (matrix(1, p, p) - diag(1, p)) + diag(1, p)
R> set.seed(1010)
R> eig <- eigen(sigma)
R> sig.sqrt <- eig$vectors %*% diag(sqrt(eig$values))
+    %*% solve(eig$vectors)
R> sig.sqrt <- (sig.sqrt + t(sig.sqrt)) / 2
R> y1 <- matrix(rnorm(n * p), nrow = n) %*% sig.sqrt
R> n <- 75
R> p <- 5
R> y2 <- matrix(rnorm(n * p), nrow = n)
R> n <- 25
R> p <- 5
R> r <- 0
R> sigma <- r * (matrix(1, p, p) - diag(1, p)) + diag(2, p)
R> y3 <- matrix(rnorm(n * p), nrow = n) %*% sqrt(sigma)
R> ycomplete <- rbind(y1, y2, y3)
R> y1[, 1] <- NA
R> y2[, c(1, 3)] <- NA
R> y3 [, 2] <- NA
R> ygroup <- rbind(y1, y2, y3)
R> out <- TestMCARNormality(data = ygroup, method = "Hawkins",
+    imputed.data = ycomplete)
R> out

Call:
TestMCARNormality(data = ygroup,
       method = "Hawkins", imputed.data = ycomplete)

Number of Patterns:  3

Total number of cases used in the analysis:  150
 Pattern(s) used:
                                 Number of cases
group.1   NA    1    1   1   1               50
group.2   NA    1   NA   1   1               75
group.3    1   NA    1   1   1               25


    Test of normality and Homoscedasticity:
  -------------------------------------------

Hawkins Test:

    P-value for the Hawkins test of normality and homoscedasticity: 1.66e-05
```

The three groups are identified, and the p value for the Hawkins test indicates heteroscedasticity. Obviously, if one does not know the distribution of the data, then the option `method = "Nonparametric"` would be appropriate to use.

Example 6: Application to real data

In this example we examine a set of data from the first wave of an on-going longitudinal study on aging (Montpetit and Bergeman 2007). These data were used by Jamshidian and Yuan (2013) to study their sensitivity analysis method to detect missing data mechanism. The data consist of 521 cases and 7 variables with 280 of the cases being complete. The variables are education, income, perceived satisfaction of social support, social coping, total life events scale, depression scale, and self-rated help. In our analysis we use 506 of the cases, as the remaining 15 cases each have a unique pattern of missingness to themselves (recall that **MissMech** requires at least two cases per missing data pattern).

```
R> data("agingdata", package = "MissMech")
R> TestMCARNormality(agingdata, del.lesscases = 1)

Call:
TestMCARNormality(data = agingdata, del.lesscases = 1)

Number of Patterns:  20

Total number of cases used in the analysis:  506

 Pattern(s) used:
          education   income   support   coping   events   depression
group.1           1        1         1        1        1            1
group.2           1       NA         1        1        1            1
group.3           1        1         1       NA        1            1
group.4           1        1         1       NA        1            1
group.5           1        1         1       NA        1           NA
group.6           1        1         1        1        1           NA
group.7           1        1        NA        1        1            1
group.8           1        1         1        1        1            1
group.9          NA        1         1        1        1            1
group.10          1        1        NA       NA        1           NA
group.11          1        1        NA       NA        1            1
group.12          1        1        NA        1        1            1
group.13         NA        1         1       NA        1            1
group.14          1        1         1        1       NA            1
group.15          1        1         1       NA       NA            1
group.16          1        1         1       NA       NA           NA
group.17         NA       NA         1       NA        1            1
group.18          1        1         1       NA       NA            1
group.19          1       NA         1       NA        1            1
group.20         NA       NA         1       NA       NA           NA
```

```
           Health   Number of cases
group.1         1               280
group.2         1                10
group.3         1                90
group.4        NA                 3
group.5         1                11
group.6         1                 6
group.7         1                14
group.8        NA                 5
group.9         1                 2
group.10       NA                 3
group.11        1                 4
group.12       NA                 2
group.13        1                 2
group.14        1                18
group.15        1                41
group.16        1                 5
group.17        1                 4
group.18       NA                 2
group.19        1                 2
group.20        1                 2


    Test of normality and Homoscedasticity:
  -------------------------------------------

Hawkins Test:

    P-value for the Hawkins test of normality and homoscedasticity: 0.000264

    Either the test of multivariate normality or homoscedasticity (or both)
    is rejected.
    Provided that normality can be assumed, the hypothesis of MCAR is
    rejected at 0.05 significance level.

Non-Parametric Test:

    P-value for the non-parametric test of homoscedasticity: 0.0164

    Hypothesis of MCAR is rejected at  0.05 significance level.
    The multivariate normality test is inconclusive.
```

There are 20 distinct patterns that include two or more cases in this data set. The Hawkins test has a very small p value, implying heteroscedasticity or non-normality. The p value for the nonparametric test is 1.6%, and thus there is sufficient evidence at 5% level that data are not MCAR. This conclusion is consistent with the findings in Jamshidian and Yuan (2013) about the missing data mechanism for these data.

4. Performance of the tests

Jamshidian and Jalal (2010) have performed a number of simulation studies to examine the performance of the Hawkins and the nonparametric tests, which we have made available in the package **MissMech**. In their study, they found that when data are normal, the Hawkins test performs well in the sense that its observed significance levels are close to the nominal significance levels, and it fails as a test of homoscedasticity and MCAR for non-normal data. The latter is due to the fact that the Hawkins test is a test of multivariate normality, in addition to homoscedasticity, and naturally has a high rejection rate for non-normal data. On the other hand, Jamshidian and Jalal (2010) found that the nonparametric test performs well for testing homoscedasticity and MCAR for both normal and non-normal data in terms of achieving observed significance levels that are close to their nominal counterparts. Moreover, they showed that both tests have reasonable powers with the Hawkins tests performing slightly better than the nonparametric test for normally distributed data. In this section we report on a small simulation study of our own that both extends and confirms findings of Jamshidian and Jalal (2010).

Table 1 shows results of a simulation study where we generated data from various symmetric and skewed distributions with various degrees of kurtosis and skewness. Each entry of the table is based on testing 1000 data sets of sample size $n = 300$ with $p = 5$ variables. In each case we generated a set of complete data, and then removed every single datum with probability of 20%, thus obtaining a data set with approximately 20% missing data that were MCAR. The middle panel of Table 1, shows the percentage of times the hypothesis of homoscedasticity (MCAR) was rejected when we set our nominal significance level at 5%. To examine how well the imputation procedure works, in each case we also performed the tests with replacing the missing data with the original generated data (i.e., using the data prior to imposition of missing) with groups being those implied by missing data pattern in each case. The result of this latter simulation is shown on the right panel of Table 1 (with the heading "Original data with no imputation"). We observe that in every case, the observed significance levels for imputed data and original data are close, which indicates that the imputation procedure is performing well.

As for the performance of the Hawkins and the nonparametric tests on the incomplete data, our results are very much in line with those reported in Jamshidian and Jalal (2010). For the normally distributed data, both the Hawkins and the nonparametric tests performed well, with the Hawkins test's observed significance level (5.8%) being closer to its nominal value of 5% as compared to the nonparametric test's observed significance level (8.7%). In our simulation study, we also generated data from the Student's t distribution with degrees of freedom ranging from 3 (heavy tailed) to 20 (close to the normal distribution). The Hawkins test's rejection rates are high for degrees of freedom ranging from 3 to 9, as expected, since data are not normally distributed. As the degrees of freedom increases, the power of the Hawkins test decreases, and in fact for $df = 20$, our rejection rate is about 11.6% since the data is close to normally distributed. The nonparametric test for the t distributed data performs well, with its rejection rates getting closer to the nominal value, as the degrees of freedom increase.

We also simulated data from a short-tailed distribution, namely Uniform$(0, 1)$. Again, as expected, the Hawkins test has a high rejection rate (95.5%) and while the observed significance level for the nonparametric test is somewhat inflated (9.2%), it is significantly better than the

	Missing data were imputed		Original data with no imputation	
Distribution	Hawkins	Nonparametric	Hawkins	Nonparametric
Normal	5.8	8.7	5.5	7.1
t, $df = 3$	100.0	10.7	100.0	10.6
t, $df = 5$	87.7	9.7	90.8	9.1
t, $df = 7$	55.1	7.2	55.9	7.9
t, $df = 9$	36.3	7.4	36.7	7.1
t, $df = 20$	11.6	6.5	11.8	6.3
Uniform$(0, 1)$	95.5	9.2	96.7	7.7
Gamma$(2, 1)$	91.1	21.0	92.1	20.3
Gamma$(5, 1)$	27.6	12.4	27.5	12.6
Gamma$(10, 1)$	12.0	9.6	12.8	9.7

Table 1: Observed significance levels for the Hawkins and nonparametric tests with $n = 300$ cases and $p = 5$ variables. The first column shows the distribution used to generate the data; results in columns 2 and 3 are based on incomplete data with MCAR missing data mechanism; results in columns 4 and 5 are based on data prior to creation of missing data.

Hawkins test. In order to see the performance of the test on skewed data, we generated data from gamma$(\alpha, \beta = 1)$. We varied the value of α from 2 (highly skewed) to 10 (approximately symmetric). As expected, the Hawkins test has a high rejection rate for the skewed data, and its rejection rates decrease as data approaches normality. The nonparametric test has a high rejection rate of 21.0% for the the highly skewed case where $(\alpha = 2)$. As the data gets closer to being symmetric, the performance of the nonparametric test improves.

What we found in our simulation, in addition to findings of Jamshidian and Jalal (2010), is that if the data are symmetrically distributed and have a high kurtosis, the nonparametric test has somewhat of an inflated rejection rate, but its performance is acceptable. On the other hand, the nonparametric test does not perform well for highly skewed data. Fortunately, however, in this case one can transform the data into symmetry and then apply the test. As we noted earlier, Jamshidian and Jalal (2010) have performed a number of power studies and we refer the reader to their paper for more information on this issue.

5. The by-product functions

While the main function in the package **MissMech** is `TestMCARNormality`, there are a total of 16 functions that are available in the package to users, each having their independent utility. In this section we describe a few of these functions that are important in their own right.

5.1. The `AndersonDarling` function

This function implements the Anderson-Darling k-sample test as described in Scholz and Stephens (1987). Given k vectors of observed values, the Anderson-Darling k-sample test tests the null hypothesis that the k samples come from a common distribution. Rejection of this test indicates that the k samples do not have the same distribution.

As an example, consider the following code, where we generate data consisting of three groups of sizes 30, 45, and 60, from normal and uniform distributions. The output shows a p value of 2.4×10^{-16}, rejecting equality of the distribution between the three groups. The remaining output values are the ingredients used to perform the test and they are described in the **MissMech** help function.

```
R> set.seed(50)
R> n1 <- 30
R> n2 <- 45
R> n3 <- 60
R> v1 <- rnorm(n1)
R> v2 <- runif(n2)
R> v3 <- rnorm(n3, 2, 3)
R> AD <- AndersonDarling(data = c(v1, v2, v3), number.cases = c(n1, n2, n3))
R> AD$pn

[1] 1.345932e-31

R> AD$adk.all

         [,1]
[1,] 6.566425
[2,] 5.075349
[3,] 6.978307

R> AD$adk

[1] 18.62008

R> AD$var.sdk

[1] 1.119502
```

5.2. The Impute function

The `Impute` function can be used as an independent tool to impute incomplete data using two different methods. The option `imputation.method = "Normal"` will impute the missing data under the assumption of the normality and the option `"Dist.Free"` uses the method of Srivastava and Dolatabadi (2009). Details of both of these methods are given in Jamshidian and Jalal (2010). Run its help function for an example.

5.3. The Mls and Ddf functions

The `Mls` function obtains the maximum likelihood estimates of mean and covariance matrix from an incomplete set of data, when it is assumed that the data come from a multivariate normal with mean $\boldsymbol{\mu}$ and covariance Σ. The EM algorithm, as described in Jamshidian and Bentler (1999) is implemented. The standard error of these estimates can be obtained by computing the negative of the Hessian of the log-likelihood, using the function `Ddf`, and taking the square root of its diagonal elements.

5.4. The `TestNey` function

The `TestNey` function uses the Neyman's smooth test of goodness of fit, as described by Ledwina (1994) to test whether a set of data come from the Uniform$(0, 1)$ distribution. An example of the use of this test is provided via its help function.

6. Discussion and other related R packages

The main applications of package **MissMech** is to test multivariate normality, homoscedasticity, and MCAR for multivariate incomplete data. The latter two tests can be performed for both normal and non-normal data. Two by-products of the **MissMech** package are testing for homoscedasticity and normality for multivariate complete data, as demonstrated in Example 5. To our knowledge, no R package exists that tests for multivariate normality or homoscedasticity in the context of incomplete data. However, for complete data, the command `mshapiro.test()` in the package **mvnormtest** (Jarek 2012) can be used to perform the Shapiro-Wilks test of multivariate normality (see e.g., Royston 1995). Also, to our knowledge, no other formal R package exists that includes test of homoscedasticity for complete data. There is a set of codes available for this purpose at `http://finzi.psych.upenn.edu/R/Rhelp02a/archive/33330.html` which uses Box's M method (Box 1949), based on the likelihood ratio test. This test is known to be highly sensitive to violations of normality, a problem that our nonparametric test deals with.

Another by-product of the **MissMech** package is that it can impute incomplete data by two methods. In one method multivariate normality of the data is assumed, and in another method, called `"Dist.Free"`, no specific distribution is assumed. There are a number of R packages that impute missing data in various contexts. For example **Amelia** II (Honaker, King, and Blackwell 2011) is a package that imputes data under the assumption of normality, **impute** (Hastie, Tibshirani, Narasimhan, and Chu 2013) has implemented the k nearest neighbor imputation, and **mi** is another more extensive package. The recent paper by Su, Gelman, Hill, and Yajima (2011) gives references to a few imputation packages and introduces the **mi** package. If one's main aim is to impute data, we recommend the **mi** package, as it has a host of methods that can handle various types of data, including categorical data, the latter being a task that **MissMech** or **Amelia** II are not designed to handle. Nonetheless, the `"Dist.Free"` method of imputation within **MissMech** package can be useful in its own right, and is not available in other R packages.

We are not aware of any R package that tests MCAR based on test of homogeneity of covariances. The R package **BaylorEdPsych** (Beaujean 2012) is the only package that we were able to find that has a test of MCAR. This test, however, is based on testing equality of means between groups with similar missing data patterns, as proposed in Little (1988). Since our package does not perform a test of equality of means to test MCAR, one might initially use the package **BaylorEdPsych** to test for MCAR based on equality of group means, and if that test is not rejected, then use our test to perform a test of MCAR for homogeneity of covariances. A word of caution, however, is that the method used in **BaylorEdPsych** assumes multivariate normality, and thus a user should perhaps use the Hawkins test in **MissMech** for multivariate normality, and if that test is not rejected use **BaylorEdPsych**. Another caution is that adjustments need to be made for multiple testing. As a future development of the **MissMech** package, we plan to add tests of MCAR based on the equality of means between groups that can handle non-normal data.

Finally, we should note that there are other statistical software that can perform imputation or test for MCAR. For example, the **EQS** software (Bentler 2006) includes tests of MCAR based on the methods proposed by Kim and Bentler (2002). However, as explained above and as noted by Jamshidian and Jalal (2010) these methods do not perform as well as the methods implemented in **MissMech**. As another example, the `MI` and `MIANALYZE` procedures in SAS create imputations and analyze the imputed data. Also `PROC DISCRIM` is a SAS (SAS Institute Inc. 2011) procedure that performs a modification of the likelihood ratio test of the homogeneity of the group covariance matrices for complete data based on the Bartlett's test (see e.g., Anderson 1984) that is well-known to be sensitive to non-normality. Unique features of the **MissMech** include (i) test of MCAR for both normal and non-normal data, (ii) test of multivariate normality for incomplete data and (iii) test of homogeneity of covariances for any combination of complete and incomplete as well as normal and non-normal data.

Acknowledgments

We would like to thank the referees for their helpful comments.

References

Anderson TW (1984). *An Introduction to Multivariate Statistical Analysis.* John Wiley & Sons, New York.

Anderson TW, Darling DA (1954). "A Test of Goodness of Fit." *Journal of the American Statistical Association*, **49**(268), 765–769.

Beaujean AA (2012). ***BaylorEdPsych**: R Package for Baylor University Educational Psychology Quantitative Courses.* R package version 0.5, URL `http://CRAN.R-project.org/package=BaylorEdPsych`.

Bentler PM (2006). ***EQS** 6 Structural Equations Program Manual.* Multivariate Software, Inc., Encino, CA.

Bentler PM, Kim KH, Yuan KH (2004). "Testing Homogeneity of Covariances with Infrequent Missing Data Patterns." Unpublished Manuscript.

Box GEP (1949). "A General Distribution Theory for a Class of Likelihood Criteria." *Biometrika*, **36**(3/4), 317–346.

Fisher RA (1932). *Statistical Methods for Research Workers.* 4th edition. Oliver and Boyd, London.

Hastie T, Tibshirani R, Narasimhan B, Chu G (2013). ***impute**: Imputation for Microarray Data.* R package version 1.36.0, URL `http://www.bioconductor.org/packages/release/bioc/html/impute.html`.

Hawkins DM (1981). "A New Test for Multivariate Normality and Homoscedasticity." *Technometrics*, **23**(1), 105–110.

Honaker J, King G, Blackwell M (2011). "**Amelia** II: A Program for Missing Data." *Journal of Statistical Software*, **45**(7), 1–47. URL http://www.jstatsoft.org/v45/i07/.

Jamshidian M, Bentler PM (1999). "ML Estimation of Mean and Covariance Structures with Missing Data Using Complete Data Routines." *Journal of Educational and Behavioral Statistics*, **24**(1), 21–41.

Jamshidian M, 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, Jansen C (2014). ***MissMech**: Testing Homoscedasticity, Multivariate Normality, and Missing Completely at Random.* R package version 1.0.1, URL http://CRAN.R-project.org/package=MissMech.

Jamshidian M, Yuan KH (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.

Jarek S (2012). ***mvnormtest**: Normality Test for Multivariate Variables.* R package version 0.1-9, URL http://CRAN.R-project.org/package=mvnormtest.

Kim KH, Bentler PM (2002). "Tests of Homogeneity of Means and Covariance Matrices for Multivariate Incomplete Data." *Psychometrika*, **67**(4), 609–623.

Ledwina T (1994). "Data-Driven Version of Neyman's Smooth Test of Fit." *Journal of the American Statistical Association*, **89**(427), 1000–1005.

Little RJA (1988). "A Test of Missing Completely at Random for Multivariate Data with Missing Values." *Journal of the American Statistical Association*, **83**(404), 1198–1202.

Montpetit A, Bergeman CS (2007). "Dimensions of Control: Mediational Analyses of the Stress-Health Relationship." *Personality and Individual Differences*, **43**(8), 2237–2248.

Neyman J (1937). "Smooth Test for Goodness of Fit." *Skandinavisk Aktuarietidskrift*, **20**, 150–199.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Royston P (1995). "A Remark on Algorithm AS 181: The W Test for Normality." *Applied Statistics*, **44**(4), 547–551.

Rubin DB (1976). "Inference and Missing Data." *Biometrika*, **63**(3), 581–592.

SAS Institute Inc (2011). *The SAS System, Version 9.3*. SAS Institute Inc., Cary, NC. URL http://www.sas.com/.

Scholz FW, Stephens MA (1987). "K-Sample Anderson-Darling Tests." *Journal of the American Statistical Association*, **82**(399), 918–924.

Srivastava MS, Dolatabadi M (2009). "Multiple Imputation and Other Resampling Schemes for Imputing Missing Observations." *Journal of Multivariate Analysis*, **100**(9), 1919–1937.

Su YS, Gelman A, Hill J, Yajima M (2011). "Multiple Imputation with Dagnostics (**mi**) in R: Opening Windows into the Black Box." *Journal of Statitical Software*, **45**(2), 315–328. URL http://www.jstatsoft.org/v45/i02/.

Wong J (2013). ***imputation**: Imputation.* R package version 2.0.1, URL http://CRAN.R-project.org/package=imputation.

Affiliation:

Mortaza Jamshidian
Department of Mathematics
California State University, Fullerton
Fullerton, CA 92834, United States of America
E-mail: mori@fullerton.edu
URL: http://math.fullerton.edu/mori

Siavash Jalal
Department of Statistics
University of California, Los Angeles
Los Angeles, CA 90095, United States of America

Camden Jansen
Department of Computer Science
University of California Irvine
Irvine, CA 92697, United States of America

Journal of Statistical Software http://www.jstatsoft.org/
published by the American Statistical Association http://www.amstat.org/

Volume 56, Issue 6 *Submitted:* 2012-01-15
January 2014 *Accepted:* 2013-06-17