Regularised Iterative Multiple Correspondence Analysis in Multiple Imputation

Submitted by

JOHANÉ NIENKEMPER

in accordance with the requirements for the

MAGISTER SCIENTIAE in MATHEMATICAL STATISTICS

(180 credits)

in the

Faculty of Natural and Agricultural Sciences

Department of Mathematical Statistics and Actuarial Science

University of the Free State

BLOEMFONTEIN

July 2013

Supervisor: Mr. M.J. von Maltitz

Declaration

I hereby declare that this dissertation, submitted for the degree M.Sc. Mathematical Statistics, at the University of the Free State, is my own independent work and has not previously been submitted, for degree purposes or otherwise, to any other institution of higher learning. I further declare that all sources cited or quoted are indicated and acknowledged by means of a comprehensive list of references. Copyright hereby cedes to the University of the Free State.

Johané Nienkemper
2007034172

21 October 2013
Date

Acknowledgements

My sincere thanks and gratitude go to the following significant influences in my life:

Heavenly Father, for graciously blessing me every day and for revealing miracles when I am weak. You are my strength.

My supervisor, Mr. MJ. von Maltitz, for his remarkable support, tireless guidance and expert advice.

My fiancé, Franré Swanepoel, for patiently waiting for me to fulfil my academic dreams. For always believing in me and showering me in his love.

My mother, Dorothy Russell, for being an active partner during this research process, from registration to submission. Her support knows no limits. This list will not suffice to describe my gratitude towards her.

My father, Johan Nienkemper, for his constant interest in my work and providing motivational notes when needed. Thank you for the support and stability provided during the transition prior to the submission of this dissertation.

My sister, Marisan Nienkemper, for always standing ready with a care-package. Thank you for your spontaneous ideas and providing laughter when it is well needed. Your confidence in me is a true support.

My extended family, Manus Conradie, for always celebrating my achievements with sincerity and a special thank you for the lending of his car during my Bloemfontein visits. Also, to Helen Swanepoel, for the support she so willingly gives, especially during the last months prior to the submission of this dissertation.

To the assessors, for their time, guidance and positive critique.

Table of Contents

List of Tables

List of Figures

Abstract

Non-responses occur commonly in survey data. The performance of a regularised iterative multiple correspondence analysis (RIMCA) algorithm in multiple imputation (MI) is compared to results obtained from single imputation (SI). RIMCA as a SI method restricts applications to data missing at random (MAR) and missing completely at random (MCAR), whereas RIMCA in MI can be adjusted to allow for missing data from the missing not at random (MNAR) mechanism as well. The RIMCA algorithm expresses multiple correspondence analysis (MCA) as a weighted principal component analysis (PCA). The success of this algorithm derives from the fact that all eigenvalues are shrunk and the last components are omitted, thus a 'double shrinkage' occurs which reduces variance and stabilises predictions. RIMCA seems to overcome overfitting and underfitting problems with regard to categorical missing data in surveys. The results obtained from simulations as well as real data are presented.

Key Terms: incomplete categorical data, missingness mechanisms, multiple imputation, multiple correspondence analysis, principal component analysis

List of Acronyms and Initialisms

AC – available-case

CA – correspondence analysis

CatPCA – categorical principal component analysis

CI – confidence interval

HNR – high percentage missing values with non-random pattern

HR – high percentage missing values with random pattern

IMCA – iterative multiple correspondence analysis

JCA – joint correspondence analysis

LD – listwise deletion

LNR – low percentage missing values with non-random pattern

LR – low percentage missing values with random pattern

MAR – missing at random

MCA – multiple correspondence analysis

MCAR – missing completely at random

MI – multiple imputation

MNAR – missing not at random

NI – non-ignorable

PCA – principal component analysis

PD – pairwise deletion

RIMCA – regularised iterative multiple correspondence analysis

RMCA – regularised multiple correspondence analysis

SI – single imputation

SVD – singular value decomposition

Definitions and Notation

Units / individuals / cases

These terms refer to the respondents of the questionnaire.

Variables

Variables refer to the questions given in the questionnaire.

Categories

Nominal scaled categorical data consists of categories of equal importance, which means that the difference between the category values cannot be determined. For this research, categories are ordinal. This means that the categories are the likert scale options available for each question. Ordinal scaled data consists of categories of different intensities or importance; therefore differences between values provide information. The ordinal scaled data enables the researcher to understand the intensity of the respondent's answer towards the specific questions.

Notation

Matrices are indicated by capital bold letters, vectors are indicated by lowercase bold letters and scalars are notated by italic letters. Notation is adapted from Rencher (2002).

Chapter 1
Introduction

"Out of Sight, Not Out of Mind"

(Buhi, Goodson & Neilands 2008)

1.1 Rationale

Incomplete data is a common occurrence in the analysis of data, in particular survey data. Missing data entries may result in a biased sample when the mechanism that causes data to become missing acts as a second round of sampling that results in a final sample that is not representative of the population in question.

The method chosen to handle the missing values in a dataset will determine the validity of results and analysis; therefore it is of utmost importance to always have the sample data reflect the population that the sample is drawn from in order to obtain accurate inferences. A range of methods exists for the handling of missing values. The most popular method is deletion (listwise deletion (LD) and pairwise deletion (PD)) (**cf**. 3.4.1.1 & 3.4.1.2). Deletion methods are the default approach in most Software packages. The procedure involves the deletion of rows in the data matrix where missing entries occur. After deletion any complete-case analysis procedure may be applied. Since the dataset will be reduced in size and in most cases will not accurately represent the population, the results produced could be biased. Therefore, deletion is an old-fashioned and inappropriate method for dealing with incomplete data.

Imputation methods are also popular, consisting of single- and multiple imputation methods. Single imputation (SI) replaces each missing value with one plausible value in order to fill the dataset to the original size (**cf**. 3.4.3.1). The most valuable imputation method is multiple imputation (MI); even with 30 years of research done in this field, researchers still attempt to develop this

procedure to the fullest. The success of MI lies in the incorporation of the uncertainty that arises from imputing missing values, therefore achieving realistic variances, whilst maintaining relationships that may occur between variables.

This study attempts to develop yet another branch of MI, investigating the applicability of a regularised iterative multiple correspondence analysis (RIMCA) algorithm to multiply impute missing values in categorical datasets.

1.1.1 Incomplete data

Missing data occurs for various reasons ranging from the capturing of data to the handling of data (**cf**. 3.3). Researchers believe that data entries become missing because of a random process, referred to as the distribution of missingness. Three missingness mechanisms can occur; missing at random (MAR), missing completely at random (MCAR) and missing not at random (MNAR). The MAR mechanism classifies missing values that are dependent on the observed values in the dataset and independent of the other missing values that occur. MCAR is an extension of the MAR mechanism, since in this case the missing values are independent of all variables in the dataset, observed and missing. Values that are missing because of the MNAR mechanism will at least be dependent on the missing values, which can also be described as the values (or questions) that were not captured by the survey (**cf**. 3.3.1).

1.2 Problem statement

This research project was inspired by an article by Josse, Chavent, Liquet and Husson (2012) on the handling of missing values by using regularised iterative multiple correspondence analysis (RIMCA) on MAR and MCAR values (Josse *et al.* 2012:93).

Josse *et al.* (2012:99) propose a RIMCA algorithm, where multiple correspondence analysis (MCA) is expressed as a weighted principal component analysis (PCA). Non-responses are imputed in questionnaire data by means of this regularised iterative MCA algorithm. The algorithm consists of three steps: initialising step, where fuzzy initial values are allocated to missing values,

reconstruction step, for the reconstruction of the indicator matrix with fuzzy entries and finally the calculation of the column margins of the new indicator matrix. This iterative process is repeated until a predetermined threshold is reached. The term fuzzy is allocated to indicator matrices consisting of matrix entries of values between zero and one (Van der Heijden & Escofier 2003:162).

Josse *et al.* (2012) experience two problems; uncertainty of what dimensions to choose, and the problem that the final fuzzy values actually have inherent uncertainties which are not modelled in the SI method. Both of these problems are solved in the adaptation of RIMCA SI to MI (**cf**. 5.5). The first problem is solved by multiply imputing over several dimensions, and the second problem by drawing multiply from the final fuzzy values when allocating categories to the imputed values. Therefore several multiple datasets are obtained for one dataset of final fuzzy values.

Research done on a RIMCA algorithm in multiple imputation (MI) has not yet been published. The RIMCA algorithm has been used as a single imputation (SI) method and performed well especially in tending to overfitting problems experienced from non-regularised algorithms. Josse *et al.* (2012:108) state that the iterative multiple correspondence analysis (IMCA) algorithm experiences difficulties with convergence and frequently converges to overfitted solutions. It has also been established by Josse *et al.* (2012:106) that the IMCA algorithm would obtain better results when applied to a perfect dataset, meaning all variables have a perfect correlation, whereas the RIMCA algorithm would always outperform IMCA in real scenarios with low correlated variables and perhaps missing values. Therefore, in this dissertation, it was decided to omit the results of the IMCA algorithm and focus on the strength of the RIMCA algorithm in MI in comparison to the RIMCA performance in SI.

As mentioned, the strength of MI lies in the fact that the uncertainty inherent in incomplete data can be incorporated into the final data analysis estimates. Rubin (2003a:620) categorises this uncertainty into three forms: *firstly*, uncertainty in the distribution of missingness; *secondly*, uncertainty in the

model and parameter values used for the imputation; and *thirdly* residual uncertainty in the drawing of the imputed values.

The measures of uncertainty will be introduced in the following way:

- Since the RIMCA algorithm is proposed for MAR and MCAR values, the missing values are considered as ignorable (**cf**. 3.3.1.4). Thus the ignorable non-responses allow the researcher to ignore the distribution of missingness (García-Laencina, Figueiras-Vidal & Sancho-Gómez 2010:266–267; Buhi *et al.* 2008:84). Therefore the distribution of missingness is not accounted for (Rubin 1978:21).

- The allocation of initial fuzzy values is done randomly in order to add additional uncertainty in the model. Another model uncertainty is introduced by not fixing the number of dimensions used in the reconstruction algorithm. The variety of dimensions will range from fuzziness (underfitting) to overfitting. All possible dimensions can be used to generate datasets; therefore the number of multiply imputed datasets will be determined by the number of dimension choices made in MI. Further, the researcher will draw multiply from the final fuzzy dataset obtained per dimension, which will result in multiple datasets of each dataset, incorporating additional model uncertainty.

- The allocation of a category value to an imputed fuzzy value will be done randomly, which will incorporate the uncertainty that is needed from actually drawing imputations – essentially this acts as the uncertainty arising from the randomness of the sampled individuals.

Thus, it is clear that all three measures of uncertainty can be met by the RIMCA algorithm, and therefore, it will be interesting to investigate the performance of this algorithm in MI. The strength of the RIMCA algorithm will be determined in the context of non-responses in questionnaire data.

1.3 Aim of the study

The aim of this study is to investigate the success of RIMCA in MI.

1.4 Objectives

- To establish whether RIMCA in MI outperforms RIMCA in SI.
- To investigate the accuracy of the predictions made by RIMCA in MI when applied to a simulated dataset.

1.5 Methodology

This quantitative research is an empirical study making use of both secondary and created data (von Maltitz 2010:15/15).

A RIMCA algorithm proposed by Josse *et al*. (2012:99) as a SI method will be applied as a MI procedure in order to compare the results obtained from the singly imputed dataset and from the multiply imputed datasets. The real dataset to be used is obtained from a satisfaction survey completed by craft operators on a waterway between two oceans located in Southern France. The dataset is referred to as *Canal des Deux Mers* and is the original dataset used by Josse *et al.* (2012:111).

A simulated dataset will be used to enable the researcher to compare a complete dataset with a multiply imputed version of the same dataset once missingness has been applied to it, in order to establish the accuracy and performance of the RIMCA algorithm in MI.

Inserting missing values in the complete simulated dataset will be done by incorporating the protocol followed by Josse *et al.* (2012:107). Two mechanisms of missingness will be considered; MCAR and MAR. In both cases two datasets will be built using a random and non-random specified pattern to insert the missing values. Also, the allowance of missing values will be determined by specified percentages. The complete discussion of the protocol will follow in Chapter Five.

1.6 Chapter outline

Chapter 1 – Introduction

This chapter describes the background for this study. The chapter is constructed by the problem statement, aim of the study, objectives and overview of the methodology. It will direct the reader to the problem of missing data, focussing on categorical questionnaire data, the different processes of missingness and the vision of the researcher to provide useful results of a RIMCA algorithm in MI.

Chapter 2 – Multivariate Techniques

A literature review on multivariate statistical techniques related to MCA is given. Dimension-reducing techniques, principal component analysis and the relationship between these methods will be discussed. Further a review on correspondence analysis and multiple correspondence analysis will be given, followed by the links between multiple correspondence analysis and other multivariate techniques. This chapter will be concluded with a discussion on a regularised version of multiple correspondence analysis.

Chapter 3 – Incomplete Data and Imputation

This chapter provides a literature review on missing values: the reason for occurrence, the type of missingness and the different approaches for the handling of missing values. The background to SI and MI is given, as well as the similarities between these approaches and the advantages and disadvantages of these techniques.

Chapter 4 – IMCA and RIMCA

The protocol proposed by Josse *et al.* (2012:97–102) for IMCA and RIMCA in SI is discussed to provide literature and background on these algorithms. The three steps of the algorithms will be shown and discussed in detail.

Chapter 5 – Methodology

This chapter will provide the background of the real data, as well as the protocol followed for the simulated data. Also, the adaptations of the algorithms discussed in the previous chapter for the implementation of the algorithms in MI will be provided. Finally, it will be shown that RIMCA satisfies the three uncertainties for MI given by Rubin (2003a:620).

Chapter 6 – Simulation Study

This chapter will argue why the use of a simulated dataset is necessary for this research project. The results obtained for the applicable objectives will be presented by means of tables and figures, which will then be discussed. Comparisons will be drawn between the results obtained from RIMCA in SI and RIMCA in MI. Further, the accuracy of the imputed values of the simulated data will be compared to the original data entries. This will enable the researcher to determine whether objective one and two of the study were achieved. Scatterplot matrices will be provided in order to establish whether the initial values allocated to missing values contribute to the final reconstructed imputed values. Further the bias, mean square errors and coverage obtained over a thousand simulations will be provided in order to compare RIMCA in SI with RIMCA in MI.

Chapter 7 – Real Categorical Dataset *Canal des Deux Mers*

The motivation for the choice of the specific dataset will be given, followed by the presentation of the results in the form of tables and figures. The chapter will be concluded by a discussion of the results. The performance of RIMCA in SI and RIMCA in MI will be compared, in order to determine whether objective one was achieved in the context of the real dataset.

Chapter 8 – Discussion and Conclusion

This chapter will discuss the results obtained from the simulation study and the real dataset, followed by the conclusions, limitations of the study and the recommendations for further research. The chapter will be concluded by focusing on the obtained results and whether the objectives and the aim of the study were met.

1.7 Summary

This introductory chapter gave a brief overview of the handling of missing data. The problem statement, aim of the study, objectives and a short description of the methodology were presented. An outline of the chapters as part of this dissertation was given.

In the following chapters the literature review of multivariate techniques and missing data approaches will be discussed, followed by a discussion on the IMCA and RIMCA algorithms. Furthermore, the methodology of the study, results obtained from an existing categorical dataset as well as a simulated dataset will be presented. This dissertation will be concluded in the discussion and conclusion chapters.

Chapter 2
Multivariate Techniques

"Social reality is multidimensional." – Pierre Bourdieu

(Le Roux & Rouanet 2004:179)

2.1 Introduction

In this chapter a literature review will be presented on dimension-reducing methods and multivariate analysis techniques and their generalisations such as: principal component analysis (PCA), correspondence analysis (CA), and multiple correspondence analysis (MCA).

2.2 Multivariate analysis

The biological and behavioural sciences were responsible for the earliest applications of multivariate techniques (Izenman 2008:2; Rencher 2002:1). The rapid development of these techniques was driven by unanswered questions in numerous fields of science and contemporary research requiring complex analysis. Most of the methods were created in the era of small- to medium-sized datasets, since analyses were constrained by the lack of powerful software programmes. Modern computers are responsible for the popularity of multivariate statistics, since they allow researchers to analyse intricate datasets (Izenman 2008:2; Rencher 2002:2; Tabachnick & Fidell 1989:1–2). An advantage of multivariate statistical analysis is to interpret the relationship between two or more related random variables, as statistical procedures are performed simultaneously on a set of random variables in order to obtain an overall result (Izenman 2008:1–2; Jackson 1991:4). According to Rencher (2002:1) the goal is to seek through the overlapping information of the correlated variables in order to obtain the underlying structure. Since simplification is the common goal of most multivariate procedures, dimension-reducing techniques play an important role.

2.3 Dimension-reducing methods

Geometrically, dimension is expressed as the rank of a matrix, which is the least amount of column and row vectors required to recreate the rows or columns of the considered matrix by means of linear combinations (Greenacre 2010:51). The rank also refers to the number of linearly independent rows of a matrix, which corresponds to the number of non-singular values for the matrix (Madsen, Hansen & Winther 2004:1). Since it is difficult to visualise and interpret data in multidimensional space, it is often useful to attempt to summarise the data as well as possible in fewer dimensions. This leads to several dimension-reduction techniques, also referred to as decomposition techniques. The decomposition of a matrix is simply a way of dividing a matrix into a set of factors, which can be orthogonal or independent. This procedure is useful in cases where the rows or columns are found to be linearly dependent, implying that the matrix in question will not be of full rank (Ientilucci 2003:1). In order to understand these techniques, it will be useful to review the decomposition methods that follow (**cf**. 2.3.1; 2.3.2 & 2.3.3).

2.3.1 Spectral decomposition

Spectral decomposition and singular value decomposition (from this point forward, SVD) are closely related dimension-reducing methods (Rencher 2002:36). Spectral decomposition expresses a real symmetric and square matrix in terms of eigenvalues and eigenvectors (Ientilucci 2003:2). The spectral decomposition of a real symmetric matrix $\mathbf{A}$ can be expressed by the following:

$$\mathbf{A} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^{\mathrm{T}},$$

where the columns of $\mathbf{U}$ represent the eigenvectors of the matrix $\mathbf{A}$ and the elements of the diagonal matrix $\boldsymbol{\Lambda}$ are the eigenvalues of matrix $\mathbf{A}$ (Madsen *et al.* 2004:2; Rencher 2002:35–36, 505).

2.3.2 Singular value decomposition

The development of SVD dates back to the 1870's and through the years has been referred to by various descriptive names: Eckart-Young decomposition, basic structure, canonical form, singular decomposition and tensor reduction. Researchers Eckart and Young were the first to apply SVD to low-rank matrix approximations in 1936, explaining the use of the Eckart-Young decomposition. Today, SVD is the common term used to refer to this dimension-reducing technique. SVD links various multivariate analysis techniques with regard to the algebra and geometry of this decomposition method. Multivariate techniques which share the relationship of SVD are: PCA, biplot analysis, CA, canonical correlation analysis and canonical variate analysis (Greenacre 1984:340–341). SVD carries great significance for dimension-reducing statistical techniques; in essence the technique breaks down a rectangular matrix into components in descending order of importance (Greenacre 2007:47).

The technique of spectral decomposition for a symmetric matrix is extended to SVD, enabling the decomposition of rectangular matrices (Ientilucci 2003:3). Therefore these methods follow a similar approach in which the eigenvalues and eigenvectors of $\mathbf{A}^{\mathrm{T}}\mathbf{A}$ (referred to as the Burt matrix) and $\mathbf{A}\mathbf{A}^{\mathrm{T}}$ are used to express the decomposition of any real matrix $\mathbf{A}$ (Rencher 2002:36, 526). The solution of the SVD enables the researcher to approximate the optimal reduced-dimension of any real matrix (Greenacre 2010:51).

The SVD of a real matrix $\mathbf{A}$ with size $n \times p$ and rank k can be expressed as:

$$\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{V}^{\mathrm{T}},$$

where $\mathbf{U}$ is $n \times k$, $\mathbf{D}$ is $k \times k$, and $\mathbf{V}$ is $p \times k$.

The elements of the diagonal matrix $\mathbf{D}$ are the non-singular values of the positive square roots of the non-zero eigenvalues of $\mathbf{A}^{\mathrm{T}}\mathbf{A}$ and $\mathbf{A}\mathbf{A}^{\mathrm{T}}$. These values are referred to as the singular values of matrix $\mathbf{A}$. The normalised eigenvectors of $\mathbf{A}\mathbf{A}^{\mathrm{T}}$ represent the k columns of $\mathbf{U}$ and the normalised eigenvectors of $\mathbf{A}^{\mathrm{T}}\mathbf{A}$ are the elements of k columns of $\mathbf{V}$. The matrices $\mathbf{U}$ and $\mathbf{V}$ are mutually orthogonal, since their columns consist of normalised eigenvectors

of symmetric matrices. This results in $\mathbf{U}^{\mathrm{T}}\mathbf{U} = \mathbf{V}^{\mathrm{T}}\mathbf{V} = \mathbb{I}$, where $\mathbb{I}$ is the identity matrix (Rencher 2002:36–37; Jolliffe 1986:37, 224; Greenacre 1984:341).

According to Wall, Rechtsteiner and Rocha (2003:91), SVD techniques are useful in three instances: a visualisation in order to express the data, representing data making use of a smaller number of variables, and detecting and extracting patterns within noisy data.

SVD is a valuable tool in the analysis of square and invertible matrices (Klema & Laub 1980:170) and is equivalent to the results obtained from diagonalisation (**cf**. 2.4.2), as well as the solution to the eigenvalue problem of the data matrix (Wall *et al.* 2003:93). Irrespective of these computational advantages, the true power of this procedure is showcased when applied to nonsquare and perhaps rank-deficient matrices (Klema & Laub 1980:170).

2.3.3 Generalised singular value decomposition

A generalisation of the definition of SVD (**cf**. 2.3.2) is given by decomposing a rectangular matrix considering constraints that may be imposed on the rows and columns of a matrix. In a standard SVD procedure a least square estimate of a given matrix by a matrix of lower rank with the same dimension will be provided, in the case of generalised singular value decomposition (GSVD) a weighted generalised least square estimate of a specific matrix will be provided. Thus in the presence of suitable constraints on the rows and columns of a matrix, the GSVD may be useful in linear multivariate techniques, such as canonical correlation and correspondence analysis (Abdi 2007:2).

In order to define GSVD consider two positive-definite square matrices, $\mathbf{D}$ of size $I \times I$ and $\mathbf{M}$ of size $J \times J$, respectively, to decompose any given matrix $\mathbf{Z}$ of size $I \times J$ (Abdi 2007:6; Greenacre 1984:344). Suppose that $\mathbf{D}$ is the matrix which expresses constraints for the rows of the matrix $\mathbf{Z}$ and $\mathbf{D}$ the constraints for the columns of the given matrix $\mathbf{Z}$. Now, the matrix $\mathbf{Z}$ can be expressed by (Abdi 2007:6):

$$\mathbf{Z} = \mathbf{C}\boldsymbol{\Lambda}\mathbf{U}^{\mathrm{T}}, \qquad with \qquad \mathbf{C}^{\mathrm{T}}\mathbf{D}\mathbf{C} = \mathbf{U}^{\mathrm{T}}\mathbf{M}\mathbf{U} = \mathbb{I}.$$

This establishes that the generalised singular vectors are only orthogonal under the limitations given by $\mathbf{D}$ and $\mathbf{M}$.

GSVD is obtained as a result of standard SVD in the following way by decomposing a given matrix $\tilde{\mathbf{Z}}$:

$$\tilde{\mathbf{Z}} = \mathbf{D}^{\frac{1}{2}}\mathbf{Z}\mathbf{M}^{\frac{1}{2}} \Leftrightarrow \mathbf{D}^{-\frac{1}{2}}\tilde{\mathbf{Z}}\mathbf{M}^{-\frac{1}{2}} = \tilde{\mathbf{Z}}.$$

Standard SVD is then performed on $\tilde{\mathbf{Z}}$:

$$\tilde{\mathbf{Z}} = \mathbf{P}\tilde{\mathbf{\Delta}}\mathbf{Q}^{\mathrm{T}} \; with \; \mathbf{P}^{\mathrm{T}}\mathbf{P} = \mathbf{Q}^{\mathrm{T}}\mathbf{Q} = \mathbb{I}.$$

Now, the matrices of the generalised eigenvectors are calculated by:

$$\mathbf{C} = \mathbf{D}^{-\frac{1}{2}}\mathbf{P} \; and \; \mathbf{V} = \mathbf{M}^{\frac{1}{2}}\mathbf{Q}.$$

The matrices of $\mathbf{\Delta}$ and $\tilde{\mathbf{\Delta}}$ containing the singular values on the diagonal are equal, thererfore:

$$\mathbf{\Delta} = \tilde{\mathbf{\Delta}}.$$

In order to verify $\mathbf{Z} = \mathbf{C}\mathbf{\Lambda}\mathbf{U}^{\mathrm{T}}$, substitution is used:

$$\begin{aligned}\mathbf{Z} &= \mathbf{D}^{-\frac{1}{2}}\tilde{\mathbf{Z}}\mathbf{M}^{-\frac{1}{2}} \\ &= \mathbf{D}^{-\frac{1}{2}}\mathbf{P}\mathbf{\Delta}\mathbf{Q}^{\mathrm{T}}\mathbf{M}^{-\frac{1}{2}} \\ &= \mathbf{C}\mathbf{\Lambda}\mathbf{U}^{\mathrm{T}}.\end{aligned}$$

2.4 Principal component analysis

The origin of PCA dates back to the work of Karl Pearson *circa* 1901. Unfortunately, its application to real datasets was stalled due to the lack of computers. PCA is a one sample technique, (where ideally no groupings occur amongst the observations in the data) which reduces the number of correlated linear variables to a set of uncorrelated transformed variables, referred to as principal components (Izenman 2008:196; Rencher 2002:380; Jackson 1991:1; Jolliffe 1986:1). In essence PCA is concerned with the associations between variables (Le Roux & Rouanet 2004:129). After the discovery of PCA a similar method was developed, referred to as factor analysis. As with PCA, factor analysis is a dimension reducing method, but factor analysis aims to explain

sets of variables using a smaller number of underlying factors (Le Roux & Rouanet 2004:130).

In PCA, the linear combinations (in the transformation) seek to account for a maximum proportion of the variance of the original variables (Rencher 2002:380). After the transformation the principal components are ordered according to the amount of variance retained, an important measure representing the amount of information provided by a specific transformed variable (Izenman 2008:196; Jackson 1991:1; Jolliffe 1986:1).

The principal component with the greatest variance is the linear combination explaining most of the data. The second principal component therefore explains second most of the data and will be geometrically illustrated in an orthogonal direction to the first principal component. Each consecutive component will be orthogonal to the prior component (Rencher 2002:380). According to Izenman (2008:196) the first principal components possessing most of the variance may be used to determine outliers, clusters of points and distributional anomalies. Izenman (2008:196) further states that principal components with a variance close to zero are considered as approximately constant, and therefore can be used to determine collinearity.

In order to summarise data as effectively as possible, the number of principal components to be retained must be accurately determined (Rencher 2002:397). A popular method is to use a 'scree plot', which is a plot of the ordered eigenvalues against their order. A visual division between large and small eigenvalues is referred to as the 'elbow' of the 'scree plot'. The order number corresponding to the component immediately before the first 'elbow' may be used as the number of principal components to be retained. The graphical technique is convenient, but lack of a definite 'elbow' may occur (Izenman 2008:205–206). Rencher (2002:397) and Jolliffe (1986:93–97) discuss other techniques such as retaining components that account for a predetermined percentage of the variance, retaining the components with eigenvalues greater than the average of the eigenvalues and, lastly, performing significance tests on the principal components responsible for the least variation.

Generally principal components are extracted from the sample covariance matrix, but in cases where the variances of specific variables are dominant or when the measurement units are dissimilar the correlation matrix may deliver more satisfying and interpretable results (Rencher 2002:383–384,393; Quinn & Keough 2002:450–451; Jackson 1991:10). The eigenvalues and eigenvectors of the covariance matrix from the PCA procedure are not easily transformed, and do not produce equivalent eigenvalues and vectors of the corresponding correlation matrix. Consequently, the principal components obtained from the covariance and correlation matrix, respectively, will not produce the same results after transformation. Since principal components based on the correlation matrix are standardised measures, they are easily compared and used in analyses, whereas principal components obtained from the covariance matrix are sensitive to the measurements of the different variables used (Jolliffe 1986:17). When the measurements of units differ greatly, results obtained from the correlation matrix will be more informative and interpretable (Jolliffe 1986:19). The term standard PCA refers to the analysis of correlations, which is PCA performed on the correlation matrix (Le Roux & Rouanet 2004:150–151, 153). Simple PCA is the analysis of covariance, which is PCA performed on the covariance matrix (Le Roux & Rouanet 2004:149–150).

Even though PCA enables the decorrelation of initial variables, reduction of dimension, and the easy identification of clusters in the data, the technique is greatly influenced by the presence of outliers (Izenman 2008:215; Jolliffe 1986:195). It must also be taken into consideration that PCA will be more effective in the presence of linear relationships between variables, since the technique makes use of association matrices (covariance or correlation matrices) (Quinn & Keough 2002:453). In order to accommodate fluctuating and nonlinear data, variations of PCA may be used and will be discussed in the following section of categorical PCA (Izenman 2008:215; Jolliffe 1986:195).

2.4.1 Categorical principal component analysis

In order to transform PCA to a nonlinear technique, researchers reformulate characteristics of the classical technique to fit the nonlinear case. This results in

a variation of categorical versions of PCA (Izenman 2008:598). As already discussed, PCA imposes linear constraints on data, consisting of assumptions that the categories of the data are ordered and constant distances between categories occur (Blasius & Greenacre 2006:30). Blasius and Greenacre's (2006:30) version of Categorical PCA (CatPCA) solves the above-mentioned distance problem and allows the distances between categories to vary, as well as taking the ordering of the categories into account. The category values of the data matrix on each dimension are replaced with optimal scale values. These optimal scale values enable the ordered categorical variables to possess nondecreasing quantification in the lower dimensions, through allowing constraints on the ordering to be imposed. This variation of PCA can be grasped as a midway between the classical linear PCA and multiple correspondence analysis (MCA), but contrary to both these methods the number of dimensions to be retained must be specified before execution (Blasius & Greenacre 2006:30).

2.4.2 Singular value decomposition in principal component analysis

The SVD dimension reduction technique is commonly considered inseparable from the multivariate technique, PCA (Greenacre 2010:59). Since SVD provides a lower rank matrix containing the least square estimates of a given matrix, maintaining the same dimension, SVD is considered equivalent to PCA and metric dimensional scaling (Abdi 2007:1). It is also recommended as the best approach to determine the principal components in PCA (Jolliffe 1986:239). The solution of PCA is provided by the result obtained from SVD, another great advantage is the format of the results of SVD, which leads to the mapping of the equivalent biplot of PCA (Greenacre 2010:59). The eigenvectors obtained from SVD are referred to as the principal components of a PCA procedure (Madsen *et al*. 2004:4). Principal components obtained from a covariance matrix (simple PCA) are equivalent to the results obtained from SVD (Wall *et al.* 2003:92–93). By centring the columns of a data matrix, $\mathbf{A}$, meaning creating zero means, the Burt matrix $\mathbf{A}^{\mathrm{T}}\mathbf{A}$ of the data matrix will be proportional to the

covariance matrix (notation **cf**. 2.3.2). The right and left singular vectors obtained from SVD are equivalent to the principal components of PCA. The singular vectors can be obtained by performing SVD, $\mathbf{A} = \mathbf{UDV}^{\mathrm{T}}$, or alternatively by the diagonalisation of the Burt matrix, $\mathbf{A}^{\mathrm{T}}\mathbf{A} = \mathbf{VD}^{2}\mathbf{V}^{\mathrm{T}}$ and then calculating $\mathbf{U} = \mathbf{AVD}^{-1}$. Also, the eigenvalues of the Burt matrix will be proportional to the principal components' variances (Wall *et al*. 2003:93)

Accroding to Jolliffe (1986:38) there are two main advantages of SVD for PCA:

- SVD is an effective method for calculating the principal components and the standardised versions of the principal components are additionally obtained.
- The SVD provides insight into what the procedure of PCA attempts to accomplish, as well as representing the results of PCA graphically and algebraically.

2.5 Correspondence analysis

The algebra within correspondence analysis (CA) can be traced back approximately 80 years, but the technique as it is known today, as the derivation of multidimensional 'scores' with a geometric interpretation, was developed around 50 years ago (Greenacre 1984:8, 11). Jean-Paul Benzécri and a small team of French data analysts studied large data tables using these methods during the early 1960's (Greenacre 1984:9). Unfortunately, the mathematical notation and style used by the French were demanding and unfamiliar; and most of their research was not translated. The only article of Benzécri translated into English was published in 1969, but since his philosophy was to focus on the data and he regarded probabilistic and mathematical modelling as irrelevant, it lacks mathematical reasoning (Greenacre 1984:9–10). After Benzécri, a series of analysts rediscovered and developed the technique (Le Roux & Rouanet 2004:23; Jackson 1991:222; Jolliffe 1986:85). A publication by Hill in 1974 which only focused on single dimensions was responsible for the popularity of CA (Greenacre 1984:11). However, Pearson would have been the founder of CA *circa* 1906 had the SVD technique been at

his disposal; this was proven by De Leeuw in a 1983 publication (Jackson 1991:223).

CA is a technique used to graphically illustrate the information in a two-way contingency table. A two-way contingency table contains the frequencies of items for a cross-classification of two categorical variables, which describes the observed association of these qualitative variables (Rencher 2002:514; Greenacre 1984:8). Points are projected onto a two-dimensional Euclidean space. The plot obtained from the two categorical variables depicts the interaction of the variables as well as the relationship between the rows and the relationship between the columns by means of a biplot (Rencher 2002:514).

A chi-square test or a log-linear model may be used to test for the significance of the associations between the categorical variables listed in the contingency table. Both these asymptotic approaches for the test of significant associations are acceptable, but the chi-square test is commonly associated with CA. In the case where insignificant associations between the two variables are found, categories in the contingency tables may be combined in order to increase specific cell frequencies. Therefore CA, is a useful tool to determine which categories should be combined, if any (Rencher 2002:515).

2.5.1 Procedure of correspondence analysis

The procedure of CA is to plot a specific point for each row and column in the contingency table, respectively. If a row point is close to a column point, this means that the combination frequency of a coordinate pair occurs more frequently. This scenario will not occur when the two variables from the contingency table were independent (Rencher 2002:515). It is expected that independent variables will produce similar row profiles, or equivalently, similar column profiles close to the origin (de Tibeiro & Murdoch 2010:519; Rencher 2002:521). The procedure of CA correlates with the coefficient of determination in linear regression, where the predictors only represent a percentage of the

possible variance and the excluded percentage is explained in the variance of the residuals, or error terms (Blasius & Greenacre 2006:8–9).

2.5.2 Objective of correspondence analysis

The objective of CA is to contract multi-dimensional data in order to explain the maximum amount of possible variation in two dimensional space. There is only a small proportion of the data that is not represented in the CA map, but it is regarded as not significantly of interest (Blasius & Greenacre 2006:8–9). The output of CA is referred to as the inertia, explained as the amount of information given by the two dimensions in the plot (Rencher 2002:515).

2.5.3 Relationship between principal component analysis and correspondence analysis

It is common to think of CA as the categorical version of PCA, concerning the geometric definition of PCA (Blasius & Greenacre 2006:5, 19). According to researchers De Leeuw and van Rijckevorsel, CA is expressed as PCA for nominal data (Jolliffe 1986:202). Both PCA and CA make use of the fact that the points of a dataset, expressed as rows and columns of a data matrix, can be displayed in a higher dimensional Euclidean space. Further, these methods aim to reduce the number of dimensions and to display the maximum variance explained by the data on preferably a two- or three-dimensional scale (Blasius & Greenacre 2006:5, 19). Both PCA and CA are procedures which focus on two aims; variable reduction and the identification of patterns in the data. Variable reduction can simply be explained by the reduction of a large set of variables to a smaller set of derived variables, which adequately represents the information provided by the data. The new derived set of variables will enable ease of execution of further analysis to be done. The patterns in the data can be revealed by making use of plots in multidimensional space, such as biplots, with regard to the new derived set of variables (Quinn & Keough 2002:443). When making use of summary variables, group structures in the data are not considered, therefore after variable reduction subsequent analyses such as graphical displays must be performed in order to obtain feasible results. PCA

and CA are concerned with the extraction of eigenvalues and eigenvectors from either correlation or covariance matrices between objects or variables (Quinn & Keough 2002:443).

2.6 Multiple correspondence analysis

CA can be extended to multiple correspondence analysis (MCA), which enables the analysis of the relationships between several categorical dependent variables (Greenacre 2010:89; Abdi & Valentin 2007:1). As already discussed, CA can be used to analyse a two-way contingency table, whereas MCA is used when the contingency table is extended to a three-way or higher-order multiway table (Rencher 2002:514, 526). As discussed in Section 2.4, the interaction of the two categorical variables, as well as the relationship between the rows and the relationship between the columns are illustrated graphically by means of a biplot (Rencher 2002:514, 526), whereas the graphical representation of MCA displays the relationships between the categories of the variables (Takane & Hwang 2006:259). Another way of expressing the purpose of both CA and MCA is as follows: CA seeks the relationship between two variables, whereas MCA is concerned with the similarities and associations within a set of two or more variables (Greenacre 2006:75).

MCA is commonly used in the visualisation of social survey data in the form of questionnaires (Blasius & Thiessen 2012:11; Josse & Husson 2012:96; Greenacre 2010:89) as a survey data screening method in which the two-dimensional map is referred to as the respondents' cognitive maps. The respondents' responses consist of answers on a discrete scale of a set of questions: "yes/no" or "strongly agree/agree/undecided/disagree/strongly disagree" (Blasius & Thiessen 2012:11; Greenacre 2010:89). Screening consists of analysing the cognitive maps by focusing on the location of the responses and respondents in search of irregularities. The data is assumed to be of high quality if there are clear patterns of confirmation or rejection obtained from the dimension with the most variation (Blasius & Thiessen 2012:11–12). Similar results will be obtained from MCA, PCA and CatPCA techniques when applied to high quality data (Blasius & Thiessen 2012:14). PCA, however, assumes input

data to have metric properties; therefore not well adapted for the screening of lower quality data (Blasius & Thiessen 2012:33).

MCA is beneficial to the analysis of multivariate categorical data (Takane & Hwang 2006:259). One MCA technique is applied by simply performing CA on a data matrix, referred to as an indicator matrix. The indicator matrix consists of a row for each subject and columns representing the response categories (Greenacre 2006:70; Takane & Hwang 2006:259; Rencher 2002:526). The number of rows is equivalent to the number of subjects and the number of columns is equivalent to the number of categories in all variables. The elements of the indicator matrix consist of ones and zeros; a one will be allocated to the corresponding category of the variable the subject has selected as a response, allocating zeros to the remaining unselected available categories in the specific row of the indicator matrix (Rencher 2002:526). The second approach in performing MCA is to perform CA on the Burt matrix (**cf**. 2.3.2) (Greenacre 2006:51, 70; Rencher 2002:526). The Burt matrix is given by $\mathbf{Z}^{\mathrm{T}}\mathbf{Z}$, where $\mathbf{Z}$ is any specific indicator matrix. Consider a dataset consisting of J individuals and a total of J categories, the Burt matrix can be expressed as follows:

$$\mathbf{Z}^{\mathrm{T}}\mathbf{Z} = \begin{bmatrix} \mathbf{Z}_1^{\mathrm{T}}\mathbf{Z}_1 & \mathbf{Z}_1^{\mathrm{T}}\mathbf{Z}_2 \dots\dots & \mathbf{Z}_1^{\mathrm{T}}\mathbf{Z}_{\mathrm{J}} \\ \mathbf{Z}_2^{\mathrm{T}}\mathbf{Z}_1 & \mathbf{Z}_2^{\mathrm{T}}\mathbf{Z}_2 \dots\dots & \vdots \\ \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots \\ \mathbf{Z}_{\mathrm{J}}^{\mathrm{T}}\mathbf{Z}_1 & \dots\dots \vdots \dots\dots & \mathbf{Z}_J^{\mathrm{T}}\mathbf{Z}_{\mathrm{J}} \end{bmatrix},$$

where the off-diagonal Burt matrix entries are two-way contingency tables which represents the associations between the sets of variables for all the individuals captured in the dataset (Greenacre 2006:50; Greenacre 1984:140).

Greenacre (2006:42) defines MCA by means of two approaches: canonical correlation analysis, which is to determine the correlation between variables, following a theoretical approach; secondly, Pearson-style principal component analysis (**cf**. 2.4), which makes use of data visualisation following a geometric approach (Greenacre 2006:43).

2.6.1 Canonical correlation analysis as MCA

Background

Canonical correlation analysis is concerned with finding the linear combinations of two subsets of variables that have maximum correlation (Rencher 2002:380; Quinn & Keough 2002:463). Therefore the technique maximises the linear correlation between the linear combinations of variables. This is done by determining the optimal scales (difference in distance between consecutive categories) between categories (Izenman 2008:223; Rencher 2002:361).

The geometry of canonical correlation analysis is responsible for the conceptualisation of the profile, mass and chi-square distance obtained in CA. Canonical correlations were therefore crucial in the theoretical development of CA. According to Greenacre (1984:108), Fisher was the first to discover the relationship in a contingency table among its optimal scaling analysis and canonical correlation analysis in 1940.

The method of canonical correlations was introduced and defined by Hotelling in 1936. Data in which the variables divide themselves into two subsets are most suited for canonical correlation analysis (Greenacre 1984:108).

Greenacre's (2006) approach to defining MCA

Firstly, two variables will be considered in order to explain the relationship between canonical correlation analysis and CA, the number of variables will then be expanded for the explanation of MCA.

Two variables

Two variables will be considered with indicator matrices given by $\mathbf{Z_1}$ and $\mathbf{Z}_2$, respectively of the same size $I \times J$, where I is the number of units and J the number of variables. The cross-product of the two indicator matrices is given by $\mathbf{Z}_1^{\mathrm{T}}\mathbf{Z}_2$, which represents the two-way contingency table of the two variables in question (**cf**. 2.6). At first an assumption is made that the scales between the categories are even, this will not be acceptable for nominal data. The scale values are contained in vectors, denoted by $\mathbf{s}_1$ and $\mathbf{s}_2$, enabling the unit

quantified responses to be contained in the vectors $\mathbf{Z}_1\mathbf{s}_1$ and $\mathbf{Z}_2\mathbf{s}_2$. Greenacre (2006:44) simplifies the notation by initial mean centring of the quantified responses, $\mathbf{1}^{\mathrm{T}}\mathbf{Z}_1\mathbf{s}_1 = \mathbf{1}^{\mathrm{T}}\mathbf{Z}_2\mathbf{s}_2 = 0$. The centred means allow for the covariance, s_{12}, between the two variables and their variances, s_1^2 and s_2^2, to be expressed by (Greenacre 2006:44):

$$s_{12} = \frac{1}{n}\mathbf{s}_1^{\mathrm{T}}\mathbf{Z}_1^{\mathrm{T}}\mathbf{Z}_2\mathbf{s}_2 = \mathbf{s}_1^{\mathrm{T}}\mathbf{P}_{12}\mathbf{s}_2$$

$$s_1^2 = \frac{1}{n}\mathbf{s}_1^{\mathrm{T}}\mathbf{Z}_1^{\mathrm{T}}\mathbf{Z}_1\mathbf{s}_1 = \mathbf{s}_1^{\mathrm{T}}\mathbf{D}_1\mathbf{s}_1 \quad \text{and} \quad s_2^2 = \frac{1}{n}\mathbf{s}_2^{\mathrm{T}}\mathbf{Z}_2^{\mathrm{T}}\mathbf{Z}_2\mathbf{s}_2 = \mathbf{s}_2^{\mathrm{T}}\mathbf{D}_2\mathbf{s}_2,$$

where $\mathbf{P}_{12} = \frac{1}{n}\mathbf{Z}_1^{\mathrm{T}}\mathbf{Z}_2$ contains the relative frequencies and is referred to as the correspondence matrix. The marginal relative frequencies, also known as the masses of the variables in question, are contained in the diagonal matrices $\mathbf{D}_1$ and $\mathbf{D}_2$ (Greenacre 2006:45).

Now the correlation can be expressed by:

$$r = \frac{s_{12}}{s_1 s_2} = \frac{\mathbf{s}_1^{\mathrm{T}}\mathbf{P}_{12}\mathbf{s}_2}{\sqrt{\mathbf{s}_1^{\mathrm{T}}\mathbf{D}_1\mathbf{s}_1\mathbf{s}_2^{\mathrm{T}}\mathbf{D}_2\mathbf{s}_2}}$$

Until now the given information solely depends on the assumption made that the scale intervals of the categories are equal. Considering the definition of canonical correlation analysis, we are concerned with the highest correlation between the variables in question. Therefore, the scale values for $\mathbf{s}_1$ and $\mathbf{s}_2$ which achieve the highest correlation must be obtained (Greenacre 2006:46). Identification conditions must be introduced that will fix the scale of the vectors $\mathbf{s}_1$ and $\mathbf{s}_2$ which will result in the highest correlation between the variables. It must be noted that the correlation will remain unchanged when any linear transformations of $\mathbf{s}_1$ and $\mathbf{s}_2$ are made. Common identification conditions are standardised variables with zero mean, $\frac{1}{n}\mathbf{1}^{\mathrm{T}}\mathbf{Z}_1\mathbf{s}_1 = \frac{1}{n}\mathbf{1}^{\mathrm{T}}\mathbf{Z}_2\mathbf{s}_2 = 0$, and a variance of one, $\mathbf{s}_1^{\mathrm{T}}\mathbf{D}_1\mathbf{s}_1 = \mathbf{s}_2^{\mathrm{T}}\mathbf{D}_2\mathbf{s}_2 = 1$ (Greenacre 2006:46). Considering these conditions, the standard coordinates of a simple CA agree fully with the optimal solution of the canonical correlation analysis (Greenacre 2006:47).

The SVD of a normalised matrix follows:

$$\mathbf{D}_1^{-\frac{1}{2}}\mathbf{P}_{12}\mathbf{D}_2^{-\frac{1}{2}} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^{\mathrm{T}}, \qquad \text{where} \qquad \mathbf{U}^{\mathrm{T}}\mathbf{U} = \mathbf{V}^{\mathrm{T}}\mathbf{V} = \mathbf{I}.$$

The singular values are contained in the diagonal matrix $\boldsymbol{\Sigma}$ and $\mathbf{U}$ is the matrix of left singular vectors as columns and $\mathbf{V}$ are the matrix containing the right singular vectors as columns.

Using the orthogonality of the singular vectors, Greenacre (2006:47) expresses one pair of left and right vectors, corresponding to a singular value of say, ϕ, by the following:

$$\mathbf{u}^{\mathrm{T}}\mathbf{D}_1^{-\frac{1}{2}}\mathbf{P}_{12}\mathbf{D}_2^{-\frac{1}{2}}\mathbf{v} = \phi.$$

The covariance formula is achieved by setting $\mathbf{s}_1 = \mathbf{D}_1^{-\frac{1}{2}}\mathbf{u}$ and $\mathbf{s}_2 = \mathbf{D}_2^{-\frac{1}{2}}\mathbf{v}$, resulting in $\mathbf{s}_1^{\mathrm{T}}\mathbf{P}_{12}\mathbf{s}_2 = \phi$. Since the identification conditions of the variance are met, the correlation is given by the singular value ϕ. Since the condition for a centred mean is not satisfied at this stage a trivial maximal solution will be obtained with a singular value equal to 1, equivalent to the vectors $\mathbf{s}_1$ and $\mathbf{s}_2$ being equal to $\mathbf{1}$. Therefore the second largest singular value of an uncentred matrix will be the maximum correlation, excluding the trivial largest singular value of 1. Centring the correspondence matrix can be easily executed as follows:

$$\mathbf{D}_1^{-\frac{1}{2}}(\mathbf{P}_{12} - \mathbf{P}_{12}\mathbf{1}\mathbf{1}^{\mathrm{T}}\mathbf{P}_{12}^{\mathrm{T}})\mathbf{D}_2^{-\frac{1}{2}} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^{\mathrm{T}}.$$

Now the first largest singular value will portray the maximum correlation. The vectors of standard coordinates on the first principal axis in CA are obtained from the solutions of $\mathbf{s}_1$ and $\mathbf{s}_2$. The largest singular value obtained, ϕ_1, is also referred to as the first canonical correlation (Greenacre 2006:48). The square roots of the principal inertias found on the axes of the map are the canonical correlations (Greenacre 2006:49). In order to determine the scale values of different sets a stepwise approach may be followed by maximising the correlation among other pairs of which the subject scores are uncorrelated with the scores already obtained as well as subject scores with differing scale values. Transforming the set of singular vectors again to standard coordinates,

now the second singular value, ϕ_2, will be the second canonical correlation. This process will continue until all canonical correlations are obtained (Greenacre 2006:48–49).

Several variables

The difference between the case of two variables and that of multiple variables is introduced by focusing on the maximisation of the correlation between the average of the two variables in question as well as the two variables themselves, whereas for the case of two variables the goal was to maximise the correlation between the two variables (Greenacre 2006:49).

In order to determine the average of two categorical variables we must first consider the indicator matrices, as specified in the two variable scenario, $\mathbf{Z}_1$ and $\mathbf{Z}_2$. The average based on the scale vectors, $\mathbf{s}_1$ and $\mathbf{s}_2$, of the quantifications of the variables are as follows (Greenacre 2006:49):

$$\frac{1}{2}(\mathbf{Z}_1\mathbf{s}_1 + \mathbf{Z}_2\mathbf{s}_2) = \frac{1}{2}[\mathbf{Z_1}\ \mathbf{Z_2}]\begin{bmatrix}\mathbf{s}_1\\ \mathbf{s}_2\end{bmatrix}$$

The combined indicator matrix of the two variables is referred to as a superindicator matrix given by $\mathbf{Z} = [\mathbf{Z}_1\ \mathbf{Z}_2]$. The separate indicator matrices have the same number of units, say n, therefore the superindicator matrix consists of $2n$ units. The correspondence matrix is notated by $\left[\frac{1}{2n}\right]\mathbf{Z}$, the matrix of row mass is $\frac{1}{n}\mathbf{I}$ and the matrix of column mass is given by $\mathbf{D} = \frac{1}{2}daig(\mathbf{D}_1, \mathbf{D}_2)$. The diagonal matrices, $\mathbf{D}_1$ and $\mathbf{D}_2$, of the two variables is combined to form the matrix $\mathbf{D}$. Considering the uncentred form of $\mathbf{Z}$, the SVD of $\mathbf{Z}$ is calculated to determine the CA solution in the following way (Greenacre 2006:49):

$$\sqrt{n}\frac{\mathbf{Z}}{2n}\mathbf{D}^{-\frac{1}{2}} = \mathbf{U\Gamma V}^{\mathrm{T}}, \qquad \text{where} \qquad \mathbf{U}^{\mathrm{T}}\mathbf{U} = \mathbf{V}^{\mathrm{T}}\mathbf{V} = \mathbf{I}.$$

Greenacre (2006:50) expresses one of the symmetric eigenvalue formulations as follows:

$$\left(\sqrt{n}\,\frac{\mathbf{Z}}{2n}\mathbf{D}^{-\frac{1}{2}}\right)^{\mathrm{T}}\left(\sqrt{n}\,\frac{\mathbf{Z}}{2n}\mathbf{D}^{-\frac{1}{2}}\right)=\frac{1}{4n}\mathbf{D}^{-\frac{1}{2}}\mathbf{Z}^{\mathrm{T}}\mathbf{Z}\mathbf{D}^{-\frac{1}{2}}=\mathbf{V}\boldsymbol{\Gamma}^{\mathbf{2}}\mathbf{V}^{\mathrm{T}}.$$

Since, $\mathbf{V}^{\mathrm{T}}\mathbf{V}=\mathbf{I}$, the above equation can be simplified to:

$$\frac{1}{4n}\mathbf{D}^{-\frac{1}{2}}\mathbf{C}\mathbf{D}^{-\frac{1}{2}}=\mathbf{V}\boldsymbol{\Gamma}^{2}\mathbf{V}^{\mathrm{T}}=\mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^{\mathrm{T}},\qquad\text{where}\qquad\mathbf{C}=\mathbf{Z}^{\mathrm{T}}\mathbf{Z}\text{ and }\boldsymbol{\Lambda}=\boldsymbol{\Gamma}^{2}.$$

The matrix $\mathbf{C}$ is indeed the Burt matrix (**cf**. 2.3.2 & 2.6), which can be expressed in terms of the two categorical variables used:

$$\mathbf{C}=\begin{bmatrix}\mathbf{Z}_1^{\mathrm{T}}\mathbf{Z}_1 & \mathbf{Z}_1^{\mathrm{T}}\mathbf{Z}_2\\ \mathbf{Z}_2^{\mathrm{T}}\mathbf{Z}_1 & \mathbf{Z}_2^{\mathrm{T}}\mathbf{Z}_2\end{bmatrix}=n\begin{bmatrix}\mathbf{D}_1 & \mathbf{P}_{12}\\ \mathbf{P}_{21} & \mathbf{D}_2\end{bmatrix}.$$

The squares of the singular values, γ^2, (and with regard to CA also referred to as the principal inertias of $\mathbf{Z}$) are contained on the diagonal of $\boldsymbol{\Gamma}^2$. Since $\boldsymbol{\Lambda}=\boldsymbol{\Gamma}^2$, it is equivalent to the diagonal matrix entries of $\boldsymbol{\Lambda}$, which is denoted by λ. When attempting to express the single symmetric eigenvalue formulation for a single eigenvector $\mathbf{v}$, $\mathbf{v}$ will be partitioned into two subvectors, $\mathbf{v}_1$ and $\mathbf{v}_2$, and the equation $\mathbf{s}=\mathbf{D}^{-\frac{1}{2}}\mathbf{v}$ must be defined. This will lead to the following eigenequation (Greenacre 2006:50):

$$\frac{1}{4}\begin{bmatrix}\mathbf{s}_1\\ \mathbf{s}_2\end{bmatrix}^{\mathrm{T}}\begin{bmatrix}\mathbf{D}_1 & \mathbf{P}_{12}\\ \mathbf{P}_{21} & \mathbf{D}_2\end{bmatrix}\begin{bmatrix}\mathbf{s}_1\\ \mathbf{s}_2\end{bmatrix}=\gamma^2=\lambda.$$

This results in:

$$\frac{1}{4}(\mathbf{s}_1^{\mathrm{T}}\mathbf{D}_1\mathbf{s}_1+\mathbf{s}_1^{\mathrm{T}}\mathbf{P}_{12}\mathbf{s}_2+\mathbf{s}_2^{\mathrm{T}}\mathbf{P}_{21}\mathbf{s}_1+\mathbf{s}_2^{\mathrm{T}}\mathbf{D}_2\mathbf{s}_2)=\gamma^2=\lambda.$$

The maximum correlation is given by $\lambda_1=\gamma_1^2$, which is the largest nontrivial eigenvalue. This is in accordance with the results obtained from simple CA with a correspondence matrix $\mathbf{P}_{12}$. The difference is that the maximum is now equal to $\frac{1}{4}(1+\phi_1+\phi_1+1)=\frac{1}{2}(1+\phi_1)$, where ϕ_1 is the canonical correlation in simple CA. The correlation between the average of the two categorical variables and either of these variables is obtained by the square of $\frac{1}{2}(1+\phi_1)$ (Greenacre 2006:50). Further, the derived versions of $\mathbf{s}_1$ and $\mathbf{s}_2$ are exactly the same as the

scale vectors from simple CA, which is the standard coordinates of the columns of the indicator matrix $\mathbf{Z}$. It is also important to take note of the eigenvalue λ_1, which is the singular value of the Burt matrix $\mathbf{C}$. Geometrically speaking with regard to CA, λ_1 is the square root of the principal inertia of the Burt matrix $\mathbf{C}$ (Greenacre 2006:51).

For the multivariable case say Q categorical variables, each with its own indicator matrix, $\mathbf{Z}_1, \mathbf{Z}_2, \ldots, \mathbf{Z}_\mathrm{Q}$ are considered. According to Greenacre (2006:51) finding a set of scale values, $\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_Q$, to maximise the overall correlation of the Q categorical variables is one of the difficulties experienced with the multivariable case. The generalisation from the two-variable case is obtained by making use of the sum of squared correlations of the scores $\mathbf{Z}_1\mathbf{s}_1, \mathbf{Z}_2\mathbf{s}_2, \ldots, \mathbf{Z}_\mathrm{Q}\mathbf{s}_Q$, with the concatenated $\mathbf{Z}_q$'s and $\mathbf{s}_q$'s forming the summation of $\mathbf{Zs}$. An overall identification constraint is introduced concerning the variance, but does not imply that the individual variances of the final solution will be equal to one. The constraint is given by, $\mathbf{s}^\mathrm{T}\mathbf{D}\mathbf{s} = 1$, where the diagonal matrix is expressed by $\frac{1}{Q} diag(\mathbf{D}_1, \mathbf{D}_2, \ldots, \mathbf{D}_\mathrm{Q})$ (Greenacre 2006:51).

The procedure to follow for the solution is similar to that of the previous cases, either CA can be performed on the superindicator matrix $\mathbf{Z} = [\mathbf{Z}_1\ \mathbf{Z}_2, \ldots, \mathbf{Z}_Q]$ or alternatively on the Burt matrix $\mathbf{C}$. In this case the Burt matrix is a block matrix with Q blocks, row-wise as well as column wise. The q^{th} categorical variable consists of J_q categories and the total number of categories is denoted by $J = \Sigma_q J_q$. The order of the superindicator matrix is $n \times J$ and the Burt matrix is of order $J \times J$. Since the row sum of $\mathbf{Z}$ is equal to a constant Q, the marginal frequencies of each categorical variable results in the column sums and the total sum of $\mathbf{Z}$ is nQ the following matrices can be specified: $\frac{1}{Qn}\mathbf{Z}$ portrays the correspondence matrix; $\frac{1}{n}\mathbf{I}$ is the row mass matrix and $\mathbf{D}$ is the column mass matrix (Greenacre 2006:51–52).

Once again the SVD of the uncentred superindicator matrix can be shown by (Greenacre 2006:52):

$$\sqrt{n}\frac{\mathbf{Z}}{Qn}\mathbf{D}^{-\frac{1}{2}} = \mathbf{U\Gamma V}^{\mathrm{T}}, \qquad \text{where} \qquad \mathbf{U}^{\mathrm{T}}\mathbf{U} = \mathbf{V}^{\mathrm{T}}\mathbf{V} = \mathbf{I}.$$

The trivial solutions can be removed by decomposing in the following manner:

$$\sqrt{n}\left(\frac{\mathbf{Z}}{Qn} - \frac{1}{n}\mathbf{1}\mathbf{1}^{\mathrm{T}}\mathbf{D}\right)\mathbf{D}^{-\frac{1}{2}},$$

where the vector of row masses is represented by $\frac{1}{n}\mathbf{1}$ and the vector of column masses by $\mathbf{1}^{\mathrm{T}}\mathbf{D}$.

Now for the second approach, performing CA on the uncentred Burt matrix $\mathbf{C}$, the SVD looks as follows (Greenacre 2006:52):

$$\mathbf{D}^{-\frac{1}{2}}\frac{\mathbf{C}}{Q^2n}\mathbf{D}^{-\frac{1}{2}} = \mathbf{V\Gamma}^2\mathbf{V}^{\mathrm{T}} = \mathbf{V\Lambda V}^{\mathrm{T}}, \qquad \text{where} \qquad \mathbf{V}^{\mathrm{T}}\mathbf{V} = \mathbf{I} \text{ and } \mathbf{C} = \mathbf{Z}^{\mathrm{T}}\mathbf{Z}.$$

The trivial solutions can be removed in a similar way by making use of the expected relative frequencies:

$$\mathbf{D}^{-\frac{1}{2}}\left(\frac{\mathbf{C}}{Q^2n} - \mathbf{D}\mathbf{1}\mathbf{1}^{\mathrm{T}}\mathbf{D}\right)\mathbf{D}^{-\frac{1}{2}}.$$

The scale values for the Q variables are given by the right-hand singular vectors, which in both approaches provide identical results. The square of the first singular value of the centred analysis of the superindicator matrix $\mathbf{Z}$ and the first singular value obtained from the centred analysis the Burt matrix $\mathbf{C}$ provides the maximum average squared correlation. Since the Burt matrix is positive definite symmetric the singular values are also the eigenvalues. The usual transformation of the singular vectors enables the calculation of the standard coordinates $\mathbf{x}$ which is partitioned into $\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_Q}$ for the Q number of variables (Greenacre 2006:52):

$$\mathbf{x} = \mathbf{D}^{-\frac{1}{2}}\mathbf{v},$$

In this case $\mathbf{v}$ is the first column of $\mathbf{V}$, which is equivalent to the first right-hand singular vector. Since differences occur between the singular values, there is a

slight difference between the principal coordinates of the two approaches (Greenacre 2006:52).

2.6.2 Pearson-style principal component analysis as multiple correspondence analysis

According to Greenacre (2006:58) the geometric approach of CA to multiple variables is difficult. In an attempt to explain the link between PCA and MCA both the chi-square distance scaling and biplot approaches will be discussed.

Background (cf**. 2.5; 2.5.1; 2.5.2 & 2.5.3**)

The map of CA is produced by performing SVD on a standardised matrix of the residuals; thereafter the principal coordinates are to be calculated. The calculation of the principal coordinates representing the points on the map are easily performed by multiplying the standard coordinates by the individual singular values. Because of the standardisation of the residuals, the standard coordinates possess unit normalisation, consequently the weighted sum of squares of the principal coordinates are normalised to be equal to the individual squared singular values of the obtained solution (Greenacre 2006:58). The principal inertia refers to the squared singular value, which corresponds to a dimension or principal axis (Greenacre 2006:59).

2.6.2.1 Chi-square distance scaling

In CA the purpose of the chi-square distance is to measure the variation between the row profiles and column profiles of a particular two-way contingency table. Considering a CA map, the difference in distance between two row points will be an optimum estimate of the chi-square distance between row profiles, equivalently the distance between two column profiles will be an ideal estimate of the chi-square distance between column profiles. The use of the chi-square distance is well known for its application in CA. Complications arise when applying this distance to the rows and columns of either a superindicator matrix $\mathbf{Z}$ or Burt matrix $\mathbf{C}$ for multivariable data. The weights used in the calculation of row profiles are obtained from the inverses of the

column masses of the correspondence matrix, the same principal applies when calculating column profiles; now making use of the row masses of the correspondence matrix as weights (Greenacre 2006:59).

Two approaches are followed when calculating intercategory relationships for column profiles (Greenacre 2006:60):

- The distances between two categories of the same variable; and
- the distances between two categories of different variables.

When using the superindicator matrix, the first approach is meaningless, since the interaction between two categories of the same variable is impossible and will result in a frequency of zero. Unfortunately both approaches for the column profile depend on marginal frequencies. Regardless of the dependency to the marginal frequencies; the second approach shows slight relevance, since the occurrence of association between two categories results in a decrease of the distance between the column points. The application of the chi-square distance on the Burt matrix performs better in comparison to the results obtained from the superindicator matrix. Calculations based on the Burt matrix are convenient, since either row or column profiles may be used, because of the Burt matrix's symmetry (Greenacre 2006:60). Both the within-variable and between-variable squared differences tend to inflate the respective distances, because of unnecessary additional terms during calculation (Greenacre 2006:60–61). Even though theoretical troubles are experienced, sufficient association patterns are allocated between variables. A contradictory result with regard to multidimensional scaling is that projections in lower dimension produce more valid results than full dimensional projections do. Also, the inflation of the total inertias of the Burt matrix and superindicator matrix leads to low percentages of inertia on the principal axes. The inflation of the distances can be explained by the dominating contribution of the diagonal matrices on the block diagonal of the Burt matrix. In conclusion, the results obtained for a two-variable dataset from both MCA and CA will produce the

same standard coordinates, but not the same principal inertias, and, consequently, not the same principal coordinates (Greenacre 2006:61).

2.6.2.2 Biplot

Background

The syllable 'bi' in biplots refers to two sets of points; these sets are rows and columns of a certain data matrix (de Tibeiro & Murdoch 2010:519; Greenacre 2010:15, 23, 24; Quinn & Keough 2002:456). The biplot is associated with data reconstruction rather than distance reconstruction (Greenacre 2006:61). A biplot is the visualisation of the rows and columns onto two- or three-dimensional planes. This visualisation is portrayed as a generalisation of a scatterplot of typically two variables. Improvement in computer software has enabled researchers to use graphical representation up to three dimensions. Since biplots hold for multi-dimensionality, the use of dimension-reducing methods play an important role in the representation of data to be illustrated in two or three dimensions (Greenacre 2010:15, 23, 24).

Fundamentally, a biplot is a point-vector plot which expresses the objects as points and the variables as lines (vectors) drawn from the origin of the plot (Quinn & Keough 2002:456).

In order to ease the visual execution of the biplot, dimension-reducing techniques (**cf**. 2.3) must be used. SVD (**cf**. 2.3.2) of the data matrix into right and left matrices provides dimensional coordinates. In practice the dimensionality is determined by the rank of the data matrix, which can accumulate to high-dimensional situations (Greenacre 2010:23–24). The SVD of a data matrix provides eigenvectors from an association matrix between variables and objects, respectively. The most general approach to construct a biplot is to use the component scores of the objects as the points and the eigenvectors of each variable relating to each component will represent the variables (Quinn & Keough 2002:456).

A great advantage of a biplot, specifically aimed at CA, is the ability to view the individuals and variables simultaneously (de Tibeiro & Murdoch 2010:519). The

biplot can be described as the optimal representation of a large number of variables which enables researchers to view the behaviour of the variables in question concurrently (Greenacre 2010:23).

In the following two sub-sections the use of the superindicator- and Burt matrices will be discussed. In both cases, the analyses of high-dimensional matrices will result in a low percentage of inertia (Greenacre 2006:64).

Superindicator matrix

Since an indicator matrix consists of zeros and ones, it is pointless to anticipate useful estimates displayed in a two-dimensional map. Other interesting measures can be used in order to make use of the biplot of an indicator matrix; the number of correct predictions made could be verified with the responses of each case (row) which is most likely to occur (Greenacre 2006:64).

Burt matrix

A similar problem occurs with the use of the Burt matrix; estimations made on the diagonal matrices of the block diagonal of the Burt matrix does not agree with the estimates obtained from the relevant contingency tables (Greenacre 2006:64). Fortunately the problem of low representation of inertia can be easily corrected by adjusting the scales of the obtained solution (Greenacre 2006:65). In order to fully grasp the technique of inertia adjustment (**cf**. 2.6.4), the procedure of joint CA will be discussed (**cf**. 2.6.3).

2.6.3 Joint CA

The problem stated is that the application of CA to the Burt matrix $\mathbf{C}$ results in the inflation of the chi-square distances as well as the total inertia being inflated by artificial amounts. These inflations occur because of the diagonal submatrices on the "diagonal" of the block Burt matrix $\mathbf{C}$. Joint CA (from this point forward, JCA) attempts to generalise simple CA in such a manner that the block diagonal matrices are ignored and only the variation in the "off-diagonal" tables are considered of the Burt matrix $\mathbf{C}$. JCA and simple CA procedures are exactly the same in the case of two-variables, which consists of one off-

diagonal table. Again different principal inertias (**cf**. 2.6.2.1) will be obtained from CA and MCA of the superindicator matrix and Burt matrix, respectively, consequently the results will not be identical to the JCA of two variables (Greenacre 2006:65).

Greenacre (2006:65–66) proposes a five step algorithm for the procedure of JCA which focuses on an alternating least-squares algorithm, treating the block diagonal matrices as missing values. The five steps follow:

- CA is applied to the Burt matrix $\mathbf{C}$ as the MCA procedure, the dimensionality of the solution must be predetermined, typically a dimension of two is selected (Greenacre 2006:65).
- In order to improve the approximation to the off-diagonal block matrices, an optional attempt is to apply adjustments to the solutions of each of the dimensions (Greenacre 2006:65).
- A reconstruction formula is used (Nenadić & Greenacre 2006:531) in order to reconstruct the values in the diagonal blocks of the Burt matrix $\mathbf{C}$. The original values in the diagonal blocks are replaced by the calculated estimates, now referring to the matrix as the modified Burt matrix $\mathbf{C}^*$ (Greenacre 2006:65).
- CA is now performed on the modified Burt matrix $\mathbf{C}^*$ (Greenacre 2006:66).
- Steps 3 and 4 are repeated until the process converges. These are the steps concerned with the replacement of the block diagonals of the Burt matrix by making use of the reconstruction formula and performing CA on the current Burt matrix. Convergence in this algorithm is obtained when a maximum absolute difference is met between the estimated values on the diagonal block of the current iteration and the corresponding estimated values on the diagonal block of the previous iteration (Greenacre 2006:66).

According to Greenacre (2006:67) there are two factors to take into consideration when calculating the percentage of inertia represented by a map:

- *Firstly*, since the dimensions are not nested in the procedure of JCA and need to be predetermined, the percentage of inertia of the two dimensions cannot be computed separately, but must be calculated together.

- *Secondly*, the usual procedure followed to determine the proportion of inertia is concerned with the ratio between the sum of the first two principal inertias and the total inertia. In the case of the modified Burt matrix, the numerator and denominator of the ratio includes the sum of the modified diagonal block values. In order to determine the percentage of inertia explained by the off-diagonal, the included contribution of the diagonal block values need to be discounted from the numerator and denominator. Nenadić and Greenacre (2006:530–533) advise on approaches to determine the additional amount obtained from the modified diagonal blocks.

2.6.4 Inertia adjustment

The main difference between MCA and JCA is a change in scale. The similarities between the two techniques enable simple scale readjustments of the MCA solution to approach the inflated percentage of inertia problem of regular MCA (Greenacre 2006:67). As discussed (**cf**. 2.6.3) the total inertia will be computed by making use of the JCA approach. This is done by removing the contributions of the diagonal blocks of the Burt matrix, which results in the adjustment of the total inertia. Another option is to calculate the average inertia of all the off-diagonal blocks directly from the tables. The following equation is used to determine the adjusted total inertia of $\mathbf{C}$ (Greenacre 2006:67):

$$average\ off-diagonal\ inertia = \frac{Q}{Q-1}\left(inertia(\mathbf{C}) - \frac{J-Q}{Q^2}\right)$$

where Q refers to the number of categorical variables and J the number of categories per categorical variable.

The principal inertias are used to calculate parts of the inertia, in the case of the superindicator matrix principal inertias are calculated by λ_s, where s represents the number of dimensions, and for the Burt matrix by λ_s^2. The adjusted inertia for each $\lambda_s \geq \frac{1}{Q}$ is calculated as follows (Greenacre 2006:68):

$$\lambda_s^{\text{adj}} = \left(\frac{Q}{Q-1}\right)^2 \left(\lambda_s - \frac{1}{Q}\right)^2$$

These principal inertias are then expressed as percentages of the average off-diagonal inertia equations given above. The obtained percentages still underestimate the inertias, when comparing to the inertias calculated by JCA, but greatly improve the solutions given by regular MCA. Greenacre (2006:68) advises that this approach should be followed when applying MCA.

2.7 Regularised MCA

The regularisation of MCA is concerned with an alternative estimation procedure that on average provides estimates that are closer to the population parameters than the conventional MCA estimation procedure would have produced. The development of regularised MCA (RMCA) evolved from ridge regression, which is able to alleviate multi-collinearity problems experienced with multiple regression analysis (Takane & Hwang 2006:260). It is expected that results obtained from a regularised procedure will be associated with smaller variance and smaller mean squared error than results from a non-regularised procedure. This shows that categories with smaller observed frequencies would perform better with the use of a regularisation parameter (Takane & Hwang 2006:277).

Incorporating prior knowledge in data analysis forms the basis of regularisation. Considering the ridge regression type of regularisation, the regularisation parameter shrinks the estimates of the category points in MCA towards the origin (Takane & Hwang 2006:278).

2.8 Conclusion

This concludes the literature review on multivariate analysis. In this chapter literature on dimension reducing techniques, PCA, as well as CA and MCA was covered. MCA was expressed by canonical correlation analysis and Pearson-style PCA, as well regularisation of MCA. The groundwork in multivariate analysis has now been laid, on which the more recent advances in multivariate analysis can be reviewed. The following chapter will consist of a literature review on the handling of missing values and imputation techniques.

Chapter 3
Incomplete data and Imputation

"The idea of imputation is both seductive and dangerous" - Dempster and Rubin (Tappen 2011:301)

3.1 Introduction

This chapter consists of a literature review on the quality of questionnaire data, incomplete data and the handling of missing values in data by making use of SI and MI methods.

3.2 Quality of data with respect to questionnaires

In order to determine the quality of data, three sources must be incorporated:

- study architecture,
- institutional agencies and
- respondent behaviour.

Study architecture refers to all elements concerning the survey design, including the question type and format of questionnaires. Institutional agencies are responsible for the distribution of questionnaires, the adequacy of interviewers and the validity of the study design. Since the data is obtained by the responses of respondents the quality of data is greatly influenced by the difficulty of the survey, the cognitive skills of the respondents as well as the interest (salience) of the topic (Blasius & Thiessen 2012:1–2, 4).

The data collection of the social and behavioural sciences mostly utilises surveys (Wang & Fan 2004:332). Two features commonly arise in survey data, which complicate the modelling process:

- Imperative restrictions, which means that certain questions will only be applicable to a subsample of the sample and would appear missing for the subsample not considered for the specific question (Raghunathan, Lepkowski, Van Hoewyk & Solenberger 2001:86).

- Certain consistency and logical bounds must be considered in the imputation process, which refers to questions with answers within specified ranges (Raghunathan *et al*. 2001:86).

Survey data mostly consist of categorical data with an ordinal or nominal scale (Blasius & Greenacre 2006:5). This implies that data collected by means of questionnaires are well suited for MCA (Josse *et al*. 2012:92).

3.3 Missing data in surveys

Data are generated or collected daily in various fields of interest. A common problem is the presence of missing data (He & Raghunathan 2009:857; Van Buuren, Brand, Groothuis-Oudshoorn & Rubin 2006:1049; Rubin 1976:581). Missing values occur due to a number of reasons, for example:

- Missing data in a social survey may arise because of negligence and loss of interest of respondents. Respondents may decide to skip questions and also refuse to provide answers to all given questions in the questionnaire (García-Laencina *et al.* 2010:264; Wang & Fan 2004:333; Pigott 2001:353; Schafer & Olsen 1998:545). Missing data may also be the result of data being lost or not properly recorded (Pigott 2001:353).

- Data collected on medical diagnoses may contain non-responses because of lack of medical equipment required for specific treatments at hospitals or treatment centres, test results might not be available immediately for preliminary analysis (García-Laencina *et al*. 2010:263–264).

- Non-responses in industrial experiments may be caused by electronic or mechanical malfunctions during the collecting of data (García-Laencina *et al*. 2010:263).

A common association made with surveys is the occurrence of non-responses (Wang & Fan 2004:332; Meng 1994:538; Rubin 1987:1), due to reasons discussed in the previous paragraph. Survey non-responses are defined by various situations; in the broadest sense non-responses include values that are missing from the processing of information given by units (individuals) rather than questions being disregarded by the respondents. An example of missing data occurring from the processing would be the occurrence of impossible values, such as weight of 700 kilograms or age exceeding 170 years. This means that values exceeding the allocated restrictions of a question are considered as missing values. Also, questions allocated for a certain subsample of the survey sample, such as income questions, will be answered by applicable respondents but not by all (Rubin 1987:1–2). The correct understanding and handling of missing data is fundamental in order to produce accurate results with smaller errors (Josse *et al.* 2012:92; von Maltitz & van der Merwe 2012:77; García-Laencina *et al*. 2010:264).

Surveys are considered as the procurement of information on a segment of a population, whereas a census of a population attempts to gain information on every unit in a population (Rubin 1987:1). The occurrences of non-responses implicate a number of concerns (Rubin 1987:1):

- The size of the database contracts and consequently less efficient estimates are obtained.
- Well known complete-case methods are not immediately viable.
- Bias can occur, which is demanding to remove since the explanation for missing values are not apparent.

Missing values are sometimes categorised with data entries that have been grouped, truncated, heaped, censored, aggregated, rounded, *etc*.; all situations

which are concerned with data losing partial information. These data editing methods are referred to as coarsened data (Schafer & Graham 2002:148; Heitjan & Rubin 1991:2244–2245). Another set of variables that are related to missing values are latent variables, which consist of unobservable or immeasurable quantities. These quantities can only be, so to say, improperly measured by questionnaires (Schafer & Graham 2002:148).

Missing data are multivariate in most practical cases, as is the data used in this research. This means that non-responses occur in multiple observed variables (von Maltitz & van der Merwe 2012:77).

'Really missing' and 'not really missing' values are the first distinction between non-responses. A value that is 'not really missing' represents a new category that conveys meaning and does not mask underlying values. Examples of such responses are: "don't know", "refuse to answer", "no preference" or "not applicable". When non-responses are categorised as 'not really missing' it is an indication that the provided questionnaire options are not sufficient for the respondent's response. Values that are 'really missing' are considered as values that should have been indicated by the individuals, and now reported as missing. These missing values mask underlying values which could contribute to the analysis of the data. 'Really missing' data can be sub-categorised into three groups explaining the reason for incomplete values which will be discussed in the following section (**cf**. 3.3.1) (Josse *et al.* 2012:92; Little & Rubin 2002:3–4).

3.3.1 Missingness mechanisms

The concept of the missingness mechanism was developed by Rubin in 1976 (Little & Rubin 2002:11). An assumption is made that values go missing because of a random process, referred to as the missingness mechanism (Kenward & Carpenter 2007:200; Little & Rubin 2002:11). The missingness mechanism is also referred to as the distribution of missingness or the response mechanism (Schafer & Graham 2002:150). The missingness of any dataset can be expressed by an indicator matrix, $\mathbf{R}$, in which missing values are represented

by the value one and observed values are depicted by zero (von Maltitz & van der Merwe 2012:77; Schafer 2003:19; Zhang 2003:582). The distribution of $\mathbf{R}$ is regarded as the missingness mechanism; the distribution is not specified, but expresses the patterns and rates of the occurring non-responses as well as illustrating the relationships between the missing values and the missingness in the data (Schafer & Graham 2002:150). Generally speaking, the missingness mechanism enables the researcher to determine whether there is a relationship between the variables and the missingness (Song & Shepperd 2007:54). The matrix $\mathbf{R}$ is treated as a set of random variables with a distribution of joint probability (Little & Rubin 2002:11; Schafer & Graham 2002:150). Before the actual handling of missing data is approached, it is important to understand the process in which the data went missing, namely the missingness mechanism (Buhi *et al.* 2008:84; Ali & Siddiqui 2000:166).

A complete data matrix, $\mathbf{Y}_{\text{comp}}$, consists of both observed values, $\mathbf{Y}_{\text{obs}}$, and missing values, $\mathbf{Y}_{\text{mis}}$. The missing values follow a distribution of positions, which are represented by the indicator matrix $\mathbf{R}$. The dimension of the complete matrix $\mathbf{Y}_{\text{comp}}$ and missingness mechanism $\mathbf{R}$, are equal (von Maltitz & van der Merwe 2012:77; Song & Shepperd 2007:54).

The three groups of non-responses that mask underlying values in the data (**cf**. 3.3) are regarded as the distributions of the missingness mechanism (Little & Rubin 2002:11-12; Rubin 1976:582, 584) and are expressed by the conditional distribution of $\mathbf{R}$ given the complete data matrix consisting of missing and observed values (García-Laencina *et al*. 2010:266). According to Buhi *et al.* (2008:84), the three mechanisms of missingness are the result of causes of missing data points, such as: bias, conditional or complete randomness and systematic explanations.

These three mechanisms are (Little & Rubin 2002:12; Schafer & Graham 2002:151):

- missing at random (MAR),
- missing completely at random (MCAR) and
- missing not at random (MNAR).

3.3.1.1 Missing at random (MAR)

MAR refers to missing values that are independent of the other missing values that occur, but may be dependent on the remaining observed variables (Little 2011:166; Little & Rubin 2002:12; Rubin 1976:582). The MAR mechanism is assumed by most missing data techniques, but is not always viable (Buhi *et al.* 2008:84; Song & Shepperd 2007:55). A simple t-test between groups with missing and complete data can be performed in order to determine whether the conditions for the MAR mechanism hold (Song & Shepperd 2007:55). MAR can be expressed by the following conditional distribution (Kenward & Carpenter 2007:201; Schafer & Graham 2002:151):

$$p\left(\mathbf{R}\middle|\mathbf{Y}_{\text{comp}}\right) = p(\mathbf{R}|\mathbf{Y}_{\text{obs}})$$

3.3.1.2 Missing completely at random (MCAR)

An extension of the MAR mechanism is MCAR, which means that the missing values and missingness are independent of all variables in the dataset, missing or observed (von Maltitz & van der Merwe 2012:77; Buhi *et al.* 2008:84; Schafer & Graham 2002:151; Little 1988:1198). The MCAR assumption does not imply that the missingness pattern is random, only that missing values are not related to the observed values (Little & Rubin 2002:12). Also MCAR implies that each unit in the dataset has the same probability of missingness and occurs by chance (Abayomi, Gelman & Levy 2008:275; Kenward & Carpenter 2007:201). The conditions for MCAR are not easily met, since this mechanism is more restrictive than MAR (Song & Shepperd 2007:55; Little & Rubin 2002:12). Since the MCAR mechanism excludes responsibilities to consider the reason for missingness, it is an ideal and extreme situation. To regulate whether MCAR is

applicable to a dataset, there should be no difference between the distributions of the observed cases and that of the missing cases. This technique is known as Little's Multivariate test. Unfortunately, Type I errors are to be expected when the sample is small (Song & Shepperd 2007:55). According to King, Honaker, Joseph and Scheve (2001:51) MAR is favoured over MCAR in many empirical situations.

The conditional distribution that expresses the MCAR mechanism is given by (Kenward & Carpenter 2007:201; Schafer & Graham 2002:151):

$$p(\mathbf{R}|\mathbf{Y}_{\text{comp}}) = p(\mathbf{R})$$

3.3.1.3 Missing not at random (MNAR)

MNAR values are dependent on the missing values in the dataset or related to values that have not been captured by the survey or data (von Maltitz & van der Merwe 2012:77-78; Abayomi *et al.* 2008:275; Little & Rubin 2002:12). The cause of missingness may be explained by systematic influences (Buhi *et al.* 2008:85). This mechanism is the worst case of missing data, since no predictions can be made from the observed values in the dataset. The problem of MNAR values cannot be solved by imputation methods or case deletion (García-Laencina *et al.* 2010:266; Song & Shepperd 2007:55). Therefore popular mechanisms and frequently used for incomplete data are MAR and MCAR (Josse *et al.* 2012:93; Jamshidian & Jalal 2010:649).

The MNAR mechanism can be expressed by the following conditional distribution (von Maltitz & van der Merwe 2012:78; Song & Shepperd 2007:55):

$$p(\mathbf{R}|\mathbf{Y}_{\text{comp}}) \neq p(\mathbf{R}|\mathbf{Y}_{\text{obs}}), \qquad p(\mathbf{R}|\mathbf{Y}_{\text{comp}})\ depends\ on\ \mathbf{Y}_{\text{mis}}$$

The following figure 3.1 illustrates the three missingness mechanisms, adapted from Schafer and Graham (2002:152):

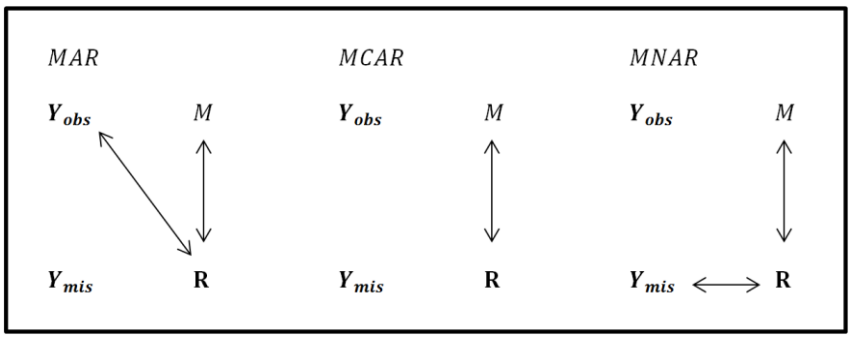


Figure 3.1 Graphical display of the missingness mechanisms

The variable M represents the causes for missing values excluding the causes of missing values in the dataset.

3.3.1.4 Ignorable and non-ignorable non-responses

The missingness mechanisms can be divided into two sub-categories (García-Laencina *et al.* 2010:267):

- Informative / non-ignorable
- Non-informative / ignorable

MAR and MCAR are classified as ignorable non-responses (Buhi *et al.* 2008:84; Schafer & Graham 2002:151; Ali & Siddiqui 2000:167) whereas MNAR is referred to as non-ignorable (Buhi *et al.* 2008:85; Song & Shepperd 2007:54; Schafer & Graham 2002:151). Ignorable non-responses enable the researcher to ignore the cause of missingness and therefore simplify the procedures for the analysis of missing data (García-Laencina *et al.* 2010:266–267; Buhi *et al*. 2008:84). The distribution of non-responses is equivalent for all the classes of the values for non-informative missing values. In contrast, the informative non-responses provide important information concerning the data (García-Laencina *et al.* 2010:266–267).

3.4 Handling of missing data

Generally three classes can be distinguished when handling missing data (Buhi *et al.* 2008:85; Song & Shepperd 2007:52; Schafer & Graham 2002:155–161):

- Deletion
- Reweighting and Toleration techniques
- Imputation

A variety of non-responses occur, thus when choosing a technique to handle the missing data the following requirements should be met (Rubin 1987:11):

- Standard complete-case methods may be used.
- Valid inferences must be produced which yield estimates that regulate for observed differences between non-respondents and respondents. Also, standard errors must be generated that mirror the abridged sample size and standard errors of the regulated observed difference between non-respondents and respondents.
- The technique must present the sensitivity of the inferences to a variety of probable models for non-responses.

3.4.1 Deletion

Deletion is the default solution for missing values in most software programmes. This outdated approach to the handling of missing data is sub-divided into listwise deletion (LD) and pairwise deletion (PD) (Buhi *et al.* 2008:85; Song & Shepperd 2007:52; Wayman 2003:3; Little & Rubin 2002:3; Schafer & Graham 2002:155). Deletion techniques are also referred to as ignoring techniques. The technique of case deletion is concerned with simply deleting cases with missing data entries (Song & Shepperd 2007:52).

A great advantage in using deletion is the ease of execution, since complete-case procedures can be applied immediately after the discarding of missing units (Buhi *et al.* 2008:85; Song & Shepperd 2007:52; Little & Rubin 2002:41;

Schafer & Graham 2002:156). According to Schafer and Graham (2002:156) however, it is a simplistic method because it performs inadequately for multivariate analyses. This is due to a number of observations that might be missing on each unit, which will result in a large fraction of the sample to be ignored (Schafer & Graham 2002:156; Schafer & Olsen 1998:546).

A distorted representation of the data may occur if the deleted values vary immensely from the remaining observed values (Penn 2007:573; Wayman 2003:2–3). Bias is to be expected when a large amount of observations are deleted (Penn 2007:576; Song & Shepperd 2007:52). Imprecise inference and bias is the result of neglecting units with missing values (Jamshidian & Jalal 2010:649). Deletion will usually produce satisfactory results under the MCAR missingness mechanism (Song & Shepperd 2007:52; Little & Rubin 2002:41; Schafer & Graham 2002:155–157; King *et al.* 2001:51). Bias may occur in non-MCAR situations, since the complete cases may not portray sufficient information on the entire population (Penn 2007:577; Little & Rubin 2002:41; Schafer & Graham 2002:157). In the presence of MAR and MNAR the observed values in general are less variable and higher than the values of the full population. This results in biases of the standard errors and the parameter estimates (Schafer & Graham 2002:157).

3.4.1.1 Listwise deletion (LD)

LD is also commonly referred to as case deletion and complete-case analysis (Buhi *et al.* 2008:85; Song & Shepperd 2007:52; Schafer & Graham 2002:155). The technique of LD is concerned with discarding each unit with at least one missing value per variable (Buhi *et al*. 2008:85–86; Schafer & Graham 2002:155).

Disadvantages of LD is the expected inflation of type II error and the loss of statistical power, since the deletion of cases leads to a smaller sample size resulting in insufficient analyses. Also, the exclusion of units from a survey causes bias, since the true population will not be represented in the data used for the analysis (Buhi *et al.* 2008:86). In the case where bias and lack of precision is small, the use of complete-case analysis will be acceptable. This will

be expected when the portion of missing data is small (Little & Rubin 2002:41–42; King *et al*. 2001:51; Schafer 1999:7).

3.4.1.2 Pairwise deletion (PD)

Complete-case analysis is not useful for univariate analyses, since values for all observations of a unit will be discarded in the presence of missing values (Little & Rubin 2002:53–54). Available-case (AC) analysis enables the use of cases where the variable of interest is available. PD is an extension of AC analysis. (Buhi *et al.* 2008:86; Song & Shepperd 2007:52). Parameters are estimated by different sets of sample units (Schafer & Graham 2002:155). Each case is evaluated separately and only variables with missing values will be discarded, thus the sample size for each unit will differ (Song & Shepperd 2007:52). Consistent results will only be achieved if the variables are weakly correlated (Pigott 2001:365). Considering regression models, PD will outperform LD when variables are moderately correlated, whereas LD performs better than PD when variables are highly correlated (Pigott 2001:363). The greatest advantage of PD is the generally higher sample size per unit than would have been the case for LD, therefore the amount of missing values are restricted (Song & Shepperd 2007:52).

3.4.2 Reweighting and toleration techniques

In the presence of non-MCAR missing data, the bias obtained from case deletion can be decreased by introducing the technique of reweighting (Little & Rubin 2002:53; Schafer & Graham 2002:157). After LD is completed the distributions of the remaining data entries are weighted in order to resemble distributions expected from a full sample. An advantage is that no models are required to determine the distribution of the population values, since weighting is a nonparametric technique (Schafer & Graham 2002:157). Reweighting is concerned with the problem of bias occurring and not with problems experienced with a fluctuating variance caused by case deletion. Therefore, this technique is most applicable in cases where the missing data is only a small

fraction of the sample size or when the sample size is large and the covariate information is restricted (Little & Rubin 2002:53).

Toleration techniques do not predict missing values, but assign probabilities to each of the values before analysis is performed on the dataset. The dataset is therefore used with non-responses. The method of toleration is useful when trying to eliminate the occurrence of bias, but unfortunately most statistical techniques require a full dataset to produce interpretable results (Song & Shepperd 2007:52).

3.4.3 Imputation

Imputation is the procedure of completing datasets by filling in plausible values for non-responses (Zhang 2003:581; Little & Rubin 2002:59; Schafer & Graham 2002:158). Imputing values will produce more reliable results than case deletion since no values will be forfeited (Schafer & Graham 2002:158). After imputation complete-case analyses may be performed on the now completed datasets (Schafer & Graham 2002:158; Raghunathan *et al*. 2001:85).

The first occurrences of imputation were concerned with replacing the missing values in a dataset by the mean or mode of the non-missing observed variables. This approach, however easily executed, was inadequate. In order to provide sound results it is essential to incorporate a degree of randomness and uncertainty when calculating confidence intervals for parameters of interest as well as standard errors (Royston 2004:228). Imputation procedures must be applied correctly; each procedure consists of its own set of assumptions and must therefore be applied with great thought. The application of *ad hoc* methods may lead to biased and misleading results, as well as the misrepresentation of the distributions of the data (Schafer & Graham 2002:147, 159; Pigott 2001:354).

A great advantage of imputation procedures is that the information provided by the observed data may be useful in the prediction of the non-responses, which enables the researcher to maintain lower variance after imputation than after deletion (Schafer & Graham 2002:158). The procedure of imputation can be

viewed as draws or means from a predictive distribution of the non-responses, which implies that a predictive distribution on the observed data must be developed for the imputation. Imputation methods allow one value to be substituted for a missing value, referred to as single imputation (**cf**. 3.4.3.1), as well as multiple values to be substituted for each missing value, known as multiple imputation (**cf**. 3.4.3.2) (Little & Rubin 2002:59).

3.4.3.1 Single imputation (SI)

Easy and popular methods when dealing with missing data are concerned with making use of complete-case analysis once datasets are filled-in (Josse & Husson 2012:80; Rubin 2003b:3; Little & Rubin 2002:85; Pigott 2001:354; Rubin 1987:11). Unfortunately the efficacy and soundness of complete-case analysis for incomplete data is not certain (Zhang 2003:581). Even though SI is easily executed, provision for uncertainty is not made. The imputed values are considered known, causing bias and the underestimation of standard errors (Penn 2007:575; Rubin 2003b:3; Little & Rubin 2002:85; Schafer & Olsen 1998:546; Rubin 1978:12–13). SI imputes a single plausible value for each missing observation in the dataset (Little & Rubin 2002:59; Pigott 2001:365). Replacing non-responses with a single value will lead to a decrease in the variance; consequently altering the distribution (Pigott 2001:365). Even if the results obtained from a single imputed dataset are accurate, it is almost certain that the amount of uncertainty from the imputer's guesswork is not captured (Meng 1994:539).

SI methods include: substitution, conditional and unconditional mean substitution, cold deck imputation, hot deck imputation or imputing from unconditional distributions and imputation from conditional distributions (von Maltitz & van der Merwe 2012:78; Little & Rubin 2002:60–62).

Substitution

Missing values are dealt with during the fieldwork stage of a survey. Units with missing values are simply replaced with units with complete responses which were not included in the sample. The dataset should however not be

considered as complete, since respondents and non-respondents may differ systematically. The substituted values must be considered as imputed values when attempting the analysis (Little & Rubin 2002:60).

Mean/mode substitution

When imputing missing values of data from a continuous distribution, the mean of the observed values will be used to replace the missing values. In the case of categorical data, the mode of the observed values will be used to replace the missing values (Ambler, Omar & Royston 2007:280).

Mean substitution differs for continuous and categorical data; in the case of continuous data, missing data will be imputed with the mean of the observed values, whereas for categorical data the mean of the observed indicator variables will be used to impute the corresponding missing indicator variables. The variance is reduced (Ambler *et al.* 2007:280) and the relationship between variables is dampened with the use of simple mean substitution (Schafer & Olsen 1998:546), this is because of the same mean value that is used to impute all the missing data entries (Wayman 2003:3).

Unconditional mean imputation

This technique follows the simple approach of mean/mode imputation except when working with categorical data. Unconditional mean imputation for categorical data is concerned with making use of the mean of the observed indicator variables to replace the corresponding missing indicator variables. The variance is reduced with the use of mean imputation (Ambler *et al.* 2007:280).

Conditional mean imputation

Conditional mean substitution is an improvement of unconditional mean substitution, since conditional means given the observed values are imputed (Little & Rubin 2002:62). Conditional mean imputation is an extension of regression imputation (Song & Shepperd 2007:52; Little & Rubin 2002:60). Predicted values obtained from a regression model based on the observed values in a dataset, replaces the missing values. Disregarding underestimated

variance, this method has been noted to perform better than MI (Song & Shepperd 2007:52).

Cold deck imputation

This process substitutes missing values with a constant value from a separate dataset than the one in question, such as a previous recorded version of the same survey. Thus after cold-decking the dataset is treated as complete, consequently complete-case techniques may be applied (Little & Rubin 2002:60–61).

Hot deck imputation

According to Little and Rubin (2002:60) this form of imputation is common in survey analyses. The method is concerned with replacing the occurring missing value of a unit with a value drawn from similar responding units in the survey (Buhi *et al.* 2008:86; Rubin 2002:60). Therefore missing values are replaced with values sampled from observed data with replacement (Ambler *et al.* 2007:280). A downfall of this technique is the assumption that respondents and nonrespondents do not differ (Buhi *et al*. 2008:86).

3.4.3.2 Multiple Imputation (MI)

Multiple imputation (MI) imputes several plausible values for a single missing data entry. Thus, after the completion of imputation the researcher has several complete datasets to analyse (Little & Rubin 2002:85; Rubin 1987:2, 15). The number of imputations is not fixed, but will be most efficient for a modest value between two and ten (Rubin 1987:2, 15). The difference between the different imputations represents the variance obtained from predicting missing values from the available observed values (Josse & Husson 2012:80). In combining the estimates from the datasets, the uncertainty is then formed from the sample variation along with the variation in the imputed values themselves (von Maltitz & van der Merwe 2012:78; Little & Rubin 2002:85). When the imputed values are derived from different models, an extra source of uncertainty concerning the correct model is included through the variation in inferences associated with the chosen models (Little & Rubin 2002:85–86). Not only is MI

a flexible approach when dealing with missing data, but also convenient (Lu, Jiang & Tsiatis 2010:1202). This procedure is compared to flexible alternative likelihood methods (Schafer 2003:22–24).

Advantages that SI and MI share are the ability to make use of complete-case methods after imputation as well as including the knowledge of the data collector (imputers' guesswork). The second advantage is greater when making use of MI, since the uncertainty incorporated for the imputer's guesswork is of two types: sampling variability given that the explanations for missing values are known and inflation in the variance due to the uncertainty about the reasons for missing values (Rubin 1978:15–16). Datasets obtained by SI techniques do not differentiate between observed and imputed values, which results in underestimated variances. MI overcomes this problem by generating distributions for the imputed and the observed values, respectively. Thus, the standard errors of the obtained estimates will incorporate the uncertainty caused by imputation (Ardington, Lam, Leibbrand & Welch 2006:826). MI enables the researcher to maintain the characteristics of the original data, such as variances and means. Imputed values should not be considered as 'guesses', but values that are allocated preserving the population variance as well as the relationships between variables (Wayman 2003:4). According to White, Wood and Royston (2007:195) another great advantage is the flexibility allowed by the method, offering analysis possibilities in the presence of missing values of all types of variables. MI has proven itself to be a robust procedure which even for small samples or large numbers of missing values can still provide satisfactory results (Wayman 2003:4).

It is important to note the three advantages of MI over SI (Rubin 1978:16):

- MI increases the accuracy of estimation, since imputations are randomly drawn when attempting to denote the distribution of the data.

- When combining the multiple datasets obtained from MI, valid inferences are attained from the repeated random draws which incorporate the additional variance caused by the non-responses in the data.

- Since MI enables imputations that are repeatedly randomly drawn under a number of models, the sensitivity of inferences to a variety of models for missing values can be studied by simply repeatedly using complete-case methods.

According to Rubin (1978:18), three disadvantages accompany MI: more work to impute multiple datasets, more storage space required for larger datasets and more analyses to be done.

Fortunately, both computer hardware and statistical software are improving at a rapid pace, providing sufficient alternatives to ease execution of such procedures (Little & Rubin 2002:86). Therefore, these disadvantages are a modest price to pay for sound inference. According to Wayman (2003:4), MI maintains a balance between simplicity and satisfying results.

MI was developed in the context of survey analyses being greatly influenced by non-responses (Reiter & Raghunathan 2007:1462; Rubin 1996:473). Even with 30 years' worth of research done in the handling of missing data, MI continues to be a growing topic with remaining unanswered questions (Cabras, Castellanos & Quirós 2011:429; Abayomi *et al.* 2008:273; Kenward & Carpenter 2007:199; Reiter & Raghunathan 2007:1462; Song & Shepperd 2007:51; King *et al.* 2001:50; Rubin 1996:473). The procedure of MI has evolved beyond non-response problems in large-sample surveys and now contributes to various settings with missing data (Reiter & Raghunathan 2007:1462; Little & Rubin 2002:85; Rubin 1996:473).

According to Rubin (1978:20), the need for MI has increased because of several factors: surveys experiencing more non-responses; the existing standard methods for missing data provide unsatisfactory results; and the handling of missing data is a growing field in statistical research.

MI is appropriate for the handling of all definitions of non-responses (**cf**. 3.3) (Rubin 1987:2). It is expedient to make use of MI, since it is a flexible method when handling missing values (Lu *et al.* 2010:1202).

The uncertainties which accompany MI complement each other in such a sense that a balance is reached producing approximately satisfactory statistical inference (Rubin 2003a:620; Zhang 2003:581). Three levels of uncertainty are taken into consideration whilst determining the correct MI process (Rubin 2003a:620; Zhang 2003:581):

- *Firstly*, uncertainty arises in choosing the distribution of the missingness mechanism;
- *Secondly*, uncertainty in the imputation model and the parameter values used to create the imputations; and
- *Thirdly*, residual uncertainty occurs when drawing imputed values.

MI accommodates the broadest definition of non-responses in surveys (Rubin 1987:2). Even though MI may be used in non-survey problems, it is especially relevant for non-responses in survey situations:

- Large amounts of data are collected in surveys, which results in a number of people tending to the data (Rubin 1996:473; Rubin 1987:3). MI is a popular approach when data is expected to be handled by a number of researchers with varying statistical skills (Raghunathan *et al.* 2001:85; Rubin 1996:473). Imprecise substitutions of missing values may occur in order to keep the database updated (Rubin 1987:3).
- As in all analyses the results obtained from surveys need to be accurate with specific estimators. It is found that complete-case methods will not provide efficient results, since the handling of missing values is done incorrectly. Also, procedures that are adapted for the presence of missing values might not be easily derived (Rubin 1987:3).

- Missing values will not commonly occur at random, as is the case for some experimental examples. An equitable assumption is that non-respondents differ from respondents, thus it is important to investigate the differences between the non-respondents and respondents and not just for the explanation behind non-responses (Rubin 1987:3–4).

- SI procedures are easily transformed and modified to enable MI (Rubin 1987:4).

MI relies on the assumption that the missingness mechanisms are ignorable (Ali & Siddiqui 2000:166). Inferences obtained from MI for MCAR and MAR are not biased (King *et al*. 2001:51).

3.5 Rubin's rules

In combining the multiply imputed datasets, Rubin's rules (1987:75–77) are followed.

Firstly, a quantity of interest (e.g. mean, variance, regression coefficient, *etc*.) is calculated from the data in question and referred to as Q. This will usually be in the form of a row vector, containing the quantity of interest for each variable in the corresponding column. In the case of complete data, inferences for Q will be based on the following (Rubin 1987:75; Rubin & Schenker 1986:367):

$$(Q - \hat{Q}) \sim N(0, U),$$

where $\hat{Q}$ estimates Q and U represents the variance of $(Q - \hat{Q})$.

Now, the first step in combining the multiple datasets is to calculate the overall average of the quantity of interest. Therefore the average of the estimates (Rubin 1987:76), which is calculated using:

$$\bar{Q}_m = \sum_{l=1}^{m} \frac{\hat{Q}_l}{m},$$

Where m is the total number of imputations and l is the current imputation.

In order to incorporate the uncertainty in $\bar{Q}$, two statistics are calculated:

$$\bar{U}_m = \sum_{l=1}^{m} \frac{\hat{U}_l}{m},$$

which calculates the average of the within-imputation variances. This can also be seen as the within-imputation variance (Schafer & Graham 2002:165).

The between-imputation variance is calculated by (Rubin & Schenker 1986:367):

$$B_m = \sum_{l=1}^{m} \frac{\left(\hat{Q}_m - \bar{Q}_m\right)^2}{m-1}$$

The total variance of $(Q - \bar{Q}_m)$ is calculated by:

$$T_m = \bar{U}_m + \left(1 + \frac{1}{m}\right) B_m.$$

Therefore the overall standard error will be equal to the square-root of T (Schafer & Graham 2002:166). If no data are missing from the data the between-imputation variance, B_m, will be equal to zero, resulting in the total variance, T_m, being equal to $\bar{U}_m$ (Schafer 1999:5; Schafer & Olsen 1998:557).

Rubin (1987:77) recommends the use of a Student's t-distribution with v degrees of freedom in order to determine confidence intervals and perform tests. When the degrees of freedom are large the t-distribution is approximately Normal. Both the between variance and average variance influence the degrees of freedom. When the between-imputation variance is larger than the average variance, the degrees of freedom will be near the value, $m-1$, where m is the number of imputations. When the average variance is larger than the between-imputation variance, very large degrees of freedom will be obtained, tending to infinity (Schafer & Olsen 1998:557). In the latter case an increase in the number of chosen imputations would not make a significant difference (Schafer & Graham 2002:166). The contrary is also true; a small v-measure indicates that an increase in the number of imputations will result in more efficient and accurate estimates with narrower confidence intervals (Schafer & Olsen 1998:557).

The degrees of freedom can be calculated by (Schafer & Graham 2002:166):

$$v = (m-1)(1+r_m^{-1})^2$$

If the missing values do not represent any information regarding the quantity of interest, Q, the imputed-data estimates will be equal and the total variance, T, will reduce to the average of the variances, $\bar{U}$. Therefore the measure, r_m, calculates the relative increase in the variance that occurs due to the missing values (Schafer 1999:5), also explained by the amount of information given by the missing values relative to the information given by the observed values (Schafer & Olsen:1998:557). This measure is calculated by:

$$r_m = \frac{\left(1+\frac{1}{m}\right)B_m}{\bar{U}_m}$$

The fraction of information about Q missing due to non-responses (rate of missing information):

$$\gamma_m = \frac{r_m + 2}{(v+3)(r_m+1)}$$

This statistic allocates the amount of information that is missing because of the non-responses, since not all of the information will be contained in the missing values, but also in the observed values (Rubin 1987:77). Also, the measure γ_m enables the researcher to understand how much more precise the estimate would have been in the presence of completely observed data (Schafer & Olsen 1998:548). Both r_m and γ_m will enable the researcher to determine the effect of the missing data on the quantity of interest, Q (Schafer & Olsen 1998:558).

The efficiency of an estimate obtained from multiple imputations in comparison to one obtained from an infinite number of imputations can be determined by:

$\left(1+\frac{\gamma}{m}\right)^{-\frac{1}{2}}$ (Rubin 1988:83) or

$\left(1+\frac{\gamma}{m}\right)^{-1}$ (Schafer 1999:7; Schafer & Olsen 1998:548),

The efficiency measure presented by Rubin (1988:83) is measured in units of standard errors, whereas Schafer (1999:7) and Schafer and Graham

(1998:548) measures in units of the variances. Rubin's efficiency measure is used in this research.

The inferences obtained from Rubin's rules are referred to as repeated-imputation inferences (Rubin 1987:75).

3.6 Methods of handling missing values in MCA

Various methods for handling missing values in MCA have been introduced and researched. In the review of Van der Heijden and Escofier (2003) the following methods are discussed and compared: missing passive, missing passive modified margin, missing single, missing multiple, missing insertion, missing fuzzy average and missing fuzzy subgroup. A popular method to use is the missing single method. This method introduces an extra category for all missing values before performing MCA on the new dataset. However convenient to use, in theory this method is appropriate when the missingness mechanism is unknown or particularly MNAR, since it assumes that missing values are related and have something in common (Van der Heijden & Escofier 2003:160, 166). The missing single method treats all missing values with regard to the same distribution of missingness, which is not suitable for values that may be missing independently of the other missing values in the dataset. Therefore missing single is not advised for MAR and MCAR values (Van der Heijden & Escofier 2003:167–168). The missing single and missing multiple methods are appropriate for data with non-random missing values (MNAR), since these methods attempt to obtain as much information from the missing values as possible. As opposed to these methods the missing passive methods, missing insertion and missing fuzzy methods do not make use of the missing values in order to obtain information (Van der Heijden & Escofier 2003:163).

The missing insertion method inserts two possibilities into the existing categories of the dataset; the most consistent responses or either the least consistent responses are inserted (Van der Heijden & Escofier 2003:161). After the insertion, a complete indicator matrix is obtained. It was found that this

method only provides adequate results for specific applications and is therefore not preferred for MCA (Van der Heijden & Escofier 2003:169).

The missing passive modified margin method developed by Escofier performs well in the presence of missing values in MCA (Josse *et al.* 2012:93; Van der Heijden & Escofier 2003:159). This method does not use row margins that are equal to the number of variables, but rather a constant row margin of $\frac{1}{n}$. This is done in order to attempt to satisfy some of the MCA properties (Van der Heijden & Escofier 2003:159). The missing passive and missing passive modified methods produce the most similar results between the various methods (Van der Heijden & Escofier 2003:164).

According to Van der Heijden and Escofier (2003:169) the missing multiple and missing insertion methods are never advisable; outliers occur with the missing multiple method and the missing insertion method is only applicable in certain circumstances. When choosing an appropriate technique to handle the missing values, the MCA properties most important for the analysis need to be considered.

3.7 Conclusion

This concludes the literature review of the occurrence of missing data and its handling. The next chapter will present literature on the IMCA and RIMCA algorithms used in this research project.

Chapter 4
IMCA and RIMCA

"Algorithm – …a sequence of computational steps that transform the input into the output."

(Cormen, Leiserson, Rivest & Stein 2001:5)

4.1 Introduction

This chapter consists of a literature review on the procedures followed by Josse *et al.* (2012) for the implementation of an iterative multiple correspondence analysis algorithm (IMCA) as well as a regularised iterative multiple correspondence analysis (RIMCA) algorithm in SI.

4.2 Background

The IMCA and RIMCA algorithms consist of three steps:

The first step is concerned with the transformation of the data matrix into an indicator matrix, followed by the allocation of initial values to missing values in the indicator matrix. Josse *et al.* (2012:97) make use of a mean imputation for continuous variables referred to as the missing fuzzy average method (**cf**. 3.6) in which missing values are substituted by the proportion observed in each category (Van der Heijden & Escofier 2003:162). In the case of this single imputation method the allocated initial values do not contribute to the total inertia and consequently do not have an effect on the outcome of analyses.

Josse *et al.* (2012:100) advise that the effect of different initial values should be explored. The only constraint when allocating initial values is the barycentric relations of the row margins per variable that should be equal to one. In CA barycentric relations are explained as the column points, also referred to as the principal coordinates, being the weighted averages of the row points, which are also referred to as the standard coordinates (Blasius & Greenacre 2006:32). Since the first step of the algorithm imputes values through the missing fuzzy

average method the relationships between individuals and variables are not taken into account.

The second step reconstructs the data and imputes plausible values to the missing values, now fuzzy values in the indicator matrix. Again fuzzy values are imputed to the missing data entries. During this step the missing values are imputed with values that are produced based on the relationships between variables and the comparisons between individuals in the dataset. Therefore these imputed values are based on the MCA axes and components, providing plausible values with respect to the observed data.

The third step involves repeating the second step until a pre-determined threshold is reached in the difference between the fuzzy values imputed from one repetition to the next, and consequently the algorithm converges.

The imputed dataset will consist of observed categorical data and imputed continuous data. The imputed values are perceived as a degree of membership to a specific category. Hence, the category with the largest proportion will be selected as the chosen category in the original dataset.

The IMCA algorithm is based on PCA, which is a continuous multivariate data technique; MCA is achieved by performing PCA on a particular triplet of variables. The discussion of the weighted PCA (**cf**. 4.2) and application of PCA on a triplet (**cf**. 4.3) will follow.

4.3 MCA as weighted PCA

There are two options when weighting any data matrix; weights may be allocated to the variables, as is the case when extracting principal components from the correlation matrix, or weights may be allocated to the observations. Weighting occurs to unify the measurements before performing analysis techniques, which may lead to interpretable and accurate results (Jackson 1991:75).

In order to illustrate MCA as a weighted PCA, a dataset with I individuals and J categorical variables $v_j, j = 1, \dots, J$ with k_j categories, is considered.

An indicator matrix of dummy variables, commonly used in MCA, is used to depict the dataset. This indicator matrix is denoted by $\mathbf{X}$ of size $I \times K$ where, $K = \sum_{j=1}^{J} k_j$.

MCA is presented as the PCA of a triplet, $(\mathbf{Z}, \mathbf{M}, \mathbf{D})$, as follows (Josse *et al.* 2012:93):

$$\left(I\mathbf{X}\mathbf{D}_{\Sigma}^{-1}, \frac{1}{IJ}\mathbf{D}_{\Sigma}, \frac{1}{I}\mathbb{I}_I\right).$$

The first term of the triplet, $\mathbf{Z}$, represents the data, the second term, $\mathbf{M}$, represents the metric and the third term, $\mathbf{D}$, represents the row masses (Josse, Chavent, Liquet & Husson 2011:4/17).

Greenacre (2010:79–80) explains the meaning of the information given in the triplet of a dataset. When referring to the PCA of a triplet, the first entry will refer to the objects/points in the multidimensional space, the second entry represents the weights and the third entry refers to the distances between them. In the case of CA or extended to MCA, the points, weights and distances are referred to as profiles, masses and chi-square distances, respectively.

The diagonal matrix of the column margins of the indicator matrix, $\mathbf{X}$, is given by $\mathbf{D}_{\Sigma} = diag\left((I_k)_{k=1,\ldots,K}\right)$. The matrix $\mathbf{M} = \frac{1}{IJ}\mathbf{D}_{\Sigma}$ is used to compute the distances between the rows. The diagonal matrix $\mathbf{D} = \frac{1}{I}\mathbb{I}_I$ corresponds to the row masses, where $\mathbb{I}_d$ is the identity matrix of size d (Josse *et al*. 2012:94).

4.4 PCA of a triplet $(\mathbf{Z}, \mathbf{M}, \mathbf{D})$

Suppose constraints are imposed on the rows and columns of given matrix; this will result in a non-symmetric distance matrix. In the case of MCA standard SVD will not suffice, therefore GSVD (**cf**. 2.3.3) should be applied. The following procedure is followed in order to perform PCA on the given triplet:

$$\mathbf{Z} = \mathbf{C}\Lambda\mathbf{U}^{\mathrm{t}},$$

where

$$\mathbf{C}^{\mathrm{t}}\mathbf{D}\mathbf{C} = \mathbf{U}^{\mathrm{t}}\mathbf{M}\mathbf{U} = \mathbb{I}_{\mathrm{r}}.$$

The matrix $\mathbf{C}$ with size $I \times r$, represents the eigenvectors of $\mathbf{ZMZ}^{\mathrm{t}}\mathbf{D}$, in descending order of the r largest eigenvalues. The matrix $\mathbf{U}$ with size $K \times r$, represents the eigenvectors of $\mathbf{Z}^{\mathrm{t}}\mathbf{DZM}$, also in descending order of the r largest eigenvalues. The rank of matrix $\mathbf{Z}$ is r. The singular values associated with the eigenvectors of $\mathbf{ZMZ}^{\mathrm{t}}\mathbf{D}$ and $\mathbf{Z}^{\mathrm{t}}\mathbf{DZM}$ are elements of the diagonal matrix, $\mathbf{\Lambda}$, which is in weakly descending order (Josse *et al*. 2012:94).

The rank of $\mathbf{Z}$ is determined by $r = K - J$, which is at most the number of non-zero eigenvalues obtained from MCA (Josse *et al*. 2012:94).

The principal components are given by $\mathbf{F} = \mathbf{C\Lambda}$, which represents the scores or the individual coordinates. This indicates that the columns of $\mathbf{C}$ correspond to the standardised principal components and the columns of $\mathbf{U}$ correspond to the axes, which automatically refers to the loadings. The first columns of the matrices $\mathbf{U}$ and $\mathbf{C}$ are equal to one and the corresponding first singular value is also equal to one, this agrees with the requirements of MCA (Josse *et al*. 2012:94).

4.5 IMCA in SI

The iterative MCA (IMCA) method uses the fact that missing values mask underlying values and therefore considers missing values from both the MCAR and MAR missingness mechanisms. The objective of the IMCA algorithm is to obtain the MCA axes and components in the presence of missing values (Josse *et al.* 2012:96).

Josse *et al.* (2012:96–97) introduces a weight matrix $\mathbf{W}$ which enables the minimisation of the reconstruction error over all non-missing values in a dataset, while ignoring the missing values. Missing values are indicated by a zero and non-missing values with a one in the proposed weight matrix, $\mathbf{W}$.

The least squares criterion is expressed as:

$$C = \| \mathbf{W} * (\mathbf{Z} - \mathbf{FU}^{\mathrm{t}}) \|^2_{\mathbf{M},\mathbf{D}},$$

with the Hadamard product, indicated by $*$.

In the presence of missing data this given criterion by Josse *et al.* (2012:97) can only be solved by iterative algorithms. The non-responses are set at initial values. When performing the analysis, the missing values are updated by repeating the procedure on the new matrix until the total change in the matrix falls below an empirically determined threshold.

As already explained (**cf**. 4.3), MCA is presented as a weighted PCA. Thus, the procedures for dealing with missing values in PCA are extended to MCA. The PCA algorithm is used, but adapted for the metric, $\mathbf{M}$, since the column margins are now dependent on the data table. This is to be expected since after the imputation of data, the column margins will change (Josse *et al*. 2012:97).

The three steps of the iterative MCA algorithm proposed by Josse *et al.* (2012:97–98) are as follows:

Take into consideration that an individual, i, has a missing value for an item, j. This will lead to a row of missing values in the indicator matrix $\mathbf{X}$ for the variable j.

STEP 1

The first step is the initialisation step ($\ell = 0$) in which the data matrix is transformed into an indicator matrix, $\mathbf{X}^0$, of dummy variables consisting of zeros and ones. The missing values are substituted by proportioned initial values, using the missing fuzzy average method (**cf**. 3.6). The proportion of the category is determined by, $\frac{I_k}{I}$. Following the procedure of Josse *et al.* (2012:97), real numbers may occur as entries of the indicator matrix satisfying the barycentric relations (Josse *et al.* 2012:99). This means that the sum of each category will be one (**cf**. 4.2).

After the substitution of the missing values, the margins of the now completed indicator matrix must be calculated. The number of variables per row will be equal to the margin of the row in particular. This implies that the k^{th} column will have a margin equal to I_k^0, which is the sum of the column entries of column k (Josse *et al*. 2012:97).

The final part of ***STEP 1*** is to calculate the diagonal matrix of the sum of the column margins given by, $\mathbf{D}_{\Sigma}^{0} = diag((I_k^0)_{k=1,\ldots,K})$.

STEP 2

The second step is three-fold; firstly, MCA is performed on the now completed data matrix, $\mathbf{X}^{\ell-1}$; this can also be seen as PCA performed on the weighted triplet (Josse *et al.* 2012:98):

$$\left(I\mathbf{X}^{\boldsymbol{\ell-1}}\left(\mathbf{D}_{\Sigma}^{\ell-1}\right)^{-1}, \frac{1}{IJ}\mathbf{D}_{\Sigma}^{\ell-1}, \frac{1}{I}\mathbb{I}_I\right).$$

The estimates $\hat{\mathbf{F}}^{\ell}$ and $\hat{\mathbf{U}}^{\boldsymbol{\ell}}$ are obtained from the singular value decomposition of $\left(I\mathbf{X}^{\ell-1}\left(\mathbf{D}_{\Sigma}^{\ell-1}\right)^{-1} - \mathbf{1}_I\mathbf{1}_K'\right) \times \sqrt{\frac{\mathbf{D}_{\Sigma}^{\ell-1}}{IJ}}$.

Secondly, a pre-determined number of dimensions, indicated by S, are chosen to be retained and will be used in the reconstruction formulae (Josse *et al.* 2011:5/17):

$$\hat{\mathbf{Z}}^{\ell} = \mathbf{1}_{\boldsymbol{I}}\mathbf{1}_{\boldsymbol{K}}' + \left(\hat{\mathbf{F}}\hat{\mathbf{U}}'\right)^{\ell}.$$

Then the associated values in the indicator matrix must be calculated using the margins of step $\ell - 1$ in order to obtain the imputed values:

$$\hat{\mathbf{X}}^{\ell} = \frac{1}{I}\hat{\mathbf{Z}}^{\ell}\mathbf{D}_{\boldsymbol{\Sigma}}^{\ell-1}.$$

Therefore, the imputed values from ***STEP 1*** will be replaced using the following update (Josse *et al.* 2011:5/17):

$$\mathbf{X}^{\ell} = \mathbf{W} * \mathbf{X} + (1 - \mathbf{W}) * \hat{\mathbf{X}}^{\ell}.$$

In the third and final phase of ***STEP 2*** the column margins, I_k^{ℓ}, of the imputed data matrix, $\mathbf{X}^{\ell}$, are calculated, resulting in the updated diagonal matrix of column margins, $\mathbf{D}_{\Sigma}^{\ell}$ (Josse *et al*. 2012:98).

STEP 3

The three phases of ***STEP 2*** are repeated until the change in the imputed indicator matrix falls below a pre-determined threshold, ε, fixed at 10^{-6}. The change is measured by $\sum_{ik}\left(\hat{x}_{ik}^{\ell-1} - \hat{x}_{ik}^{\ell}\right)^2 \leq \varepsilon$ (Josse *et al*. 2012:98). The final

dataset is obtained by replacing the fuzzy imputed values with category values. This is done per variable for each observation; the category with the largest fuzzy value will be allocated the chosen category, therefore allocating the most plausible category values with respect to a degree of membership (Josse *et al*. 2012:99).

Discussion

One disadvantage of this method is that the missing values are imputed in such a way that they do not contribute to the total inertia (variance) of the data. Therefore even with imputation the missing values are essentially 'skipped'. This method is also referred to as the 'reconstruction of order 0' which is used in CA (Josse *et al*. 2012:98).

Overfitting is the main problem experienced with IMCA (Josse *et al*. 2012:100). This happens when the value of the least squares criterion concerning the weight matrix of the iterative MCA is low. A low criterion value could represent good fit for the observed values, but not for the predicted values, which could occur because of poor estimation of the axes and components. In the MCA context overfitting results in points which represent individuals and categories located far from each other. The structure of the dataset, as well as the number of missing values and dimensions kept for the reconstruction may result in overfitting (Josse *et al*. 2012:101).

Methods to reduce overfitting include the reduction of the number of dimensions for the imputation step, which however must be done selectively to secure the amount of information obtained from the dimensions. Another approach is to use shrinkage methods. Similar to the ridge estimator in regression, a regularised iterative MCA algorithm is introduced to overcome the high variance obtained from the prediction of missing values (Josse *et al*. 2012:101).

4.5.1 RIMCA in SI

The regularised version of the IMCA algorithm discussed in the previous section (**cf**. 4.5) stabilises the predictions obtained from the IMCA algorithm that

according to Josse *et al.* (2012:101–102) obtained high variance for the estimation of the axes and components as well as the prediction of the non-responses.

The three steps of the regularised IMCA algorithm proposed by Josse *et al.* (2012:102) are as follows:

Step 1

In the context of SI, the IMCA and RIMCA algorithms follow the same procedure for the first step (**cf**. 4.5).

Step 2

During the second step of the RIMCA algorithm a regularisation of the reconstruction formulae is introduced which stabilises the variance. A 'shrinkage' term is added to the IMCA reconstruction formulae. This 'shrunk'-variance reconstruction step is based on an algorithm to perform PCA on an incomplete dataset proposed by Josse in 2009 (Josse *et al.* 2012:102).

The reconstruction formulae (**cf**. 4.5) can be rewritten with λ_s representing the eigenvalue of rank s, which is also equal to the variance of each component $\mathbf{f}_s$, as follows (Josse *et al.* 2011:9/17):

$$\hat{\mathrm{z}}_{ik}{}^{\ell} = 1 + \sum_{s=1}^{S} \frac{\hat{f}_{is}^{\ell}}{\| \hat{\mathbf{f}}_s^{\ell} \|} \left(\sqrt{\lambda_s}\right) \hat{u}_{ks}^{\ell} \, .$$

Subsequently this step is replaced by:

$$\hat{\mathrm{z}}_{ik}{}^{\ell} = 1 + \sum_{s=1}^{S} \frac{\hat{f}_{is}^{\ell}}{\| \hat{\mathbf{f}}_s^{\ell} \|} \left(\sqrt{\lambda_s} - \frac{\hat{\sigma}^2}{\sqrt{\lambda_s}} \right) \hat{u}_{ks}^{\ell} \, ,$$

where σ^2 is estimated by the mean of the last eigenvalues:

$$\hat{\sigma}^2 = \frac{1}{K-J-S} \sum_{s=S+1}^{K-J} \lambda_s .$$

The indicator matrix is updated following the same procedure as for the IMCA algorithm, firstly allocating the associated values and then replacing the fuzzy missing average method imputed values with the values obtained from the reconstruction steps (**cf**. 4.5).

Step 3

The third step remains unchanged from the IMCA algorithm.

Discussion

The motivation of the RIMCA algorithm is to remove the error (noise) to consequently increase stability in prediction. The assumption is made that the last dimensions only contain noise, whereas both information and noise are contained in the first dimensions. This explains why the noise parameter is estimated by the mean of the last eigenvalues (or variances of the last principal components). In the extreme case where no overfitting occurs in the presence of missing values, the RIMCA and the IMCA algorithms will achieve the same results. When the noise has a great effect on the analysis, the regularisation will shrink the coordinates of the individuals towards the average (Josse *et al*. 2012:103).

The success of this algorithm derives in the fact that all eigenvalues are shrunk and the last components are omitted, thus the RIMCA algorithm allows for a 'double shrinkage' (Josse *et al*. 2012:103).

4.6 Conclusion

This concludes the literature and theory on the procedures followed for the IMCA and RIMCA algorithms in SI. The methodology of this research and the adaptations of the algorithms in SI for the implementation of the RIMCA algorithm in MI will be presented in the following chapter.

Chapter 5
Methodology

"The more abstract the truth you want to teach, the more you will have to win over the senses in its favor" – Nietzsche

(Le Roux & Rouanet 2004:1)

5.1 Introduction

This chapter presents of the research design and methodology followed for the study objectives. This will be followed by the research population, as well as the specification of the simulation protocol and the description of the real data. The difference between the algorithms in SI and MI will be discussed with regard to the uncertainties required for MI, given by Rubin (2003a:620 & **cf**. 1.2).

5.2 Research design

A quantitative research approach is followed for this empirical research project (von Maltitz 2010:15/15). Both secondary (existing) data and simulated (created) data are used in order to investigate the performance of the algorithms in a real data scenario as well as the strength of the imputations obtained from the algorithms in comparison to the complete simulated data. The researcher has low control over the existing data, but high control over the simulated data, since the data is created by the researcher using a pre-determined protocol. The existing and simulated data are numerical categorical datasets. The real dataset consists of observed data and missing data entries, whereas the simulated dataset is complete before the missing entries are allocated through specified random and non-random patterns.

5.3 Objectives

The table 5.1 demonstrates the various procedures that are used in each of the objectives. Since the objectives are based on comparisons, simultaneous columns are selected per objective.

Table 5.1 Procedures used for the objectives

Objective	RIMCA	SI	MI
1	✓	✓	✓
2	✓	✓	✓

5.3.1 Objective one: To establish whether RIMCA in MI outperforms RIMCA in SI

This objective is attained by applying the RIMCA algorithm as a SI technique to the real and simulated datasets with observed data and non-responses, as well as applying the RIMCA algorithm as a MI technique to the same datasets. The means and confidence intervals obtained from both RIMCA procedures are compared in order to determine whether RIMCA performs better in SI or MI. Rubin's rules (**cf**. 3.5) will be used for the calculation of the descriptive statistics obtained by MI, the confidence intervals for the means of the singly imputed datasets will be determined by use of the Student's t-distribution (Rice 1995:182):

$$\frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t_{n-1}$$

95% Confidence intervals ($t_{0.025}(n-1)$ and $t_{0.075}(n-1)$) will be obtained and used in the analysis.

5.3.2 Objective two: To investigate the accuracy of the predictions made by RIMCA in MI when applied to a simulated dataset

The performance of the RIMCA algorithm in MI is determined by comparing the original completed simulated datasets, with low and high correlation structure

respectively, with the imputed datasets. This is done through an apparent error rate:

$$error\ rate = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{(x - \hat{x}) \div 2}{IJ}$$

The error rate will enable the researcher to evaluate the performance of the predictions made by the RIMCA algorithm for each of the $S = K - J$ dimensions, where I is the number of individuals and J the total number of variables and x is an indicator matrix of the original data, whereas $\hat{x}$ is the imputed corresponding indicator matrix. An error rate is calculated for each of the possible S dimensions in each of the simulated datasets. These simulated datasets consist of different missingness mechanisms, percentages of missing values and random or non-random patterns (**cf**. 5.4.1.1). Therefore the amount of apparent error rates will be determined by the number of generated datasets as well as the possible dimensions available for the datasets in question. The apparent error rates of RIMCA in SI will also be calculated, in order to determine the difference in accuracy of both procedures.

5.4 Study population

Simulated datasets and an existing dataset are used. Table 5.2 illustrates the data used for each objective.

Table 5.2 Data allocation to objectives

Objective	Simulated data	Real data
1	✓	✓
2	✓	

5.4.1 Simulated data

The simulated dataset represents the responses of 100 individuals to ten questions (variables) with three possible categories per variable. Missing values are inserted into the simulated dataset by following random and non-random patterns, as well as incorporating MCAR and MAR missingness mechanisms. The protocol used for the simulation will be discussed in the following section.

5.4.1.1 Simulation protocol

The protocol followed by Josse *et al.* (2012:107) is used for the simulation of the data.

The simulated datasets consist of 100 individuals and ten variables which are drawn from a multivariate Normal distribution. The correlations for the covariance matrices are fixed at 0.4 for the low structure and 0.8 for the strong structure. A two-block diagonal covariance matrix is simulated; one block of size 6 x 6 and the remaining block of size 4 x 4. The same level of correlation is used for both blocks of the diagonal covariance matrix. The ten variables are each distributed in three equal-count categories.

Two missingness mechanisms are taken into account; MCAR and MAR, inserted with random and non-random patterns.

The MCAR case includes 10% and 30% missing values at random over all the variables or following a specific non-random pattern. The non-random pattern inserts missing values for the first three variables of the first 60 individuals and missing values are allocated for the ninth and tenth variables of the last 60 individuals. Since these proposed non-random patterns for missing values are independent of any variables in the data, the MCAR mechanism is well represented.

The MAR case includes 8% and 16% missing values, again either at random, excluding variable one and seven, or by inserting a non-random pattern. The non-random pattern consists of missing values in the second to sixth variables when the first variable allocates the first category and for the eighth to tenth variables when the seventh variable allocates the last category.

Considering the construction of the simulated datasets, the number of underlying dimensions will be four. This is achieved since a simulated dataset consists of two independent blocks of variables, each containing three possible categories per variable. This results in two underlying dimensions for each block. Josse *et al.* (2012:107) states that the use of a cross-validation algorithm also suggested four underlying dimensions.

Since the researcher aims to evaluate the performance of RIMCA in MI, the number of dimensions will not be fixed for the purpose of this research. Not fixing the number of dimensions will increase the uncertainty in the model, which will incorporate one of the uncertainties required by MI. This will be discussed in more detail in the following section (**cf**. 5.5).

Table 5.3 provides a summary of the given protocol information.

Table 5.3 Summary of simulation protocol

	MCAR	MAR
Correlation Level	Low structure: 0.4 High structure: 0.8	Low structure: 0.4 High structure: 0.8
Percentage of missing values	10% and 30%	8% and 16%
Random pattern	Random	Random
Non-random pattern	Variables 1-3 for individuals 1 to 40 Variables 1-3, 9 and 10 for individuals 41 to 100	Variables 2-6 when variable 1 allocated the first category Variables 8-10 when variable 7 allocated the third category

Therefore eight different sets of data per missingness mechanism (MCAR and MAR) are simulated. These eight datasets consist of two covariance matrices of low and high correlation structure, allowing for two pre-determined percentages of missing values for each level of correlation. The percentages of missing values are either inserted at random or by following a non-random pattern. This results in eight simulated datasets per missingness mechanism. Therefore a total of 16 simulated datasets.

5.4.2 Real data

The real dataset originated from a user satisfaction survey of craft operators on the *Canal des Deux Mers*, in the South of France. The survey was performed by *Voes Navigables de France*, a public corporation responsible for the development of the largest network of navigable waterways in Europe.

The questionnaire consists of the responses of 1232 individuals to 14 questions with two or three possible categories, with a total of 35 categories. Approximately 9% (9.0677%) of the values in the dataset are missing and non-

responses occur in 42.5% of the respondents. The 14 questions (variables) of the *Canal des Deux Mers* survey are presented in Appendix I.

5.5 From SI to MI

The procedures followed for the IMCA and RIMCA algorithms in SI were discussed in the preceding chapter (**cf**. 4.5 & 4.5.1). The adjustments made to satisfy the measures of uncertainty of MI will now be discussed.

Uncertainty in the distribution of the missingness mechanism

As was discussed in the first chapter of this dissertation (**cf**. 1.2), the RIMCA algorithm is proposed for MAR and MCAR values, the missing values are considered as ignorable (**cf**. 3.3.1.4). Thus the ignorable non-responses allow the researcher to ignore the distribution of missingness (García-Laencina, Figueiras-Vidal & Sancho-Gómez 2010:266–267; Buhi *et al.* 2008:84). Therefore the distribution of missingness is not accounted for (Rubin 1978:21).

Uncertainty in the parameter values used for the imputations

This uncertainty is concerned with the uncertainty of the model as well as the uncertainty in the parameters used for the imputations. This is achieved by three changes from SI to MI.

- In the case of the SI procedures the initial values are allocated proportionally with regard to the observed number of categories in the dataset, $\frac{I_k}{I}$ (**cf**. 4.5). Therefore, additional uncertainty in the model will be incorporated by allowing for random starting points. Therefore the substitution of the missing values with initial values in the indicator matrix is not done by making use of the missing fuzzy average method as was the case for SI (***STEP 1*** of the SI algorithms (**cf**. 4.5)); instead randomly generated Uniform(0,1) initial values are allocated for the category value of a particular variable, placing a constraint over the category values per variable to add up to one, in order to satisfy the barycentric relations required for MCA (**cf**. 4.2 & 4.5).

- The number of dimensions to retain in the reconstruction algorithm *a priori* will not be fixed, in contrast to the procedure by Josse *et al.* (2012:100). This decision is based on the convergence of the different dimensions: it is found that on average all the dimensions converge after the same reasonable number of iterations of the algorithms. The choice of such a small threshold used ($\varepsilon = 10^{-6}$) in the third step of the IMCA and RIMCA algorithm (**cf**. 4.5 & 4.5.1) can be explained by the fast convergence reached by the algorithms. (Tables providing the average number of iterations of the RIMCA algorithm before it converges are given in Appendix J.) Therefore, all possible dimensions, $S = K - J - 1$, can be used in order to generate imputed datasets in the case of RIMCA and $S = K - J$ dimensions in the case of IMCA. Thus a range of datasets are obtained of which some would possibly be underfitting (fuzziness) or overfitting. This incorporates the uncertainty in the model.

The number of multiple datasets to use in MI is recommended to be a modest number between two and ten (Rubin 1987:2, 15). In this research ten multiple datasets are randomly chosen from the possible S dimensions that are available. Therefore the multiple datasets represent imputations made upon a randomly selected dimension. The ten dimensions used to generate ten imputed datasets are selected by the random integer generation function in MATLAB (*randi*). In order to determine whether the results obtained from RIMCA are sensitive towards a random selection of dimensions, a type of sensitivity analysis is performed. This consists of the repeating the process of the selection of dimensions ten times, followed by the imputation of these multiple datasets and their analysis, using Rubin's rules (**cf**. 3.5). The sensitivity analysis will only be done for the simulated data. Each repetition will consist of a newly simulated dataset from which ten newly selected random dimensions will be used to obtain multiple datasets, which will then be combined using Rubin's rules. The ten repetitions will be displayed graphically in order to show the stability across different dimensions for each variable (**cf**. 6.3). This will

enable the researcher to determine whether the results obtained from the multiply imputed datasets fluctuate significantly over a selection of randomly appointed dimensions.

- Five datasets will be drawn randomly from the final fuzzy indicator matrix obtained after the convergence of RIMCA, as explained in the following uncertainty section. Therefore, multiple datasets for each multiple dataset (determined by a specific dimension) will be generated, which will incorporate additional model uncertainty. Thus the between-variation (**cf**. 3.5) will capture uncertainty about the model and the imputation.

Uncertainty when drawing imputed values

The uncertainty in the drawing of imputations is achieved in the transformation of the final fuzzy reconstructed values to a matrix of categorical data. The SI procedures allocate category values to the largest fuzzy value per variable, thus allocating with regard to a degree of membership. In order to obtain uncertainty in the actual imputations, the category values for an imputed data point will be allocated at random, based on a randomly generated Uniform(0,1) value. Thus, the values remaining for each category of a missing data point would indicate the probability of that category being drawn as an imputed value for that missing data point. Since the final fuzzy values are fixed after the algorithm converges, the parameters of the imputation model are also fixed. This means that imputations are drawn conditionally on estimates of the parameters, therefore the MI variance might be underestimated.

Table 5.4 tabulates the differences between SI and MI with respect to the three steps of the algorithms (**cf**. 4.5 & 4.5.1).

Table 5.4 Differences between SI and MI

	SI	MI
Step 1	Expected proportions for initial values	Random initial values
Step 2	Fixed number of dimensions	Allows for a range of dimensions
Step 3	Repeat Step 2 until a pre-determined threshold is reached. For final dataset: allocate category value by means of a degree of membership	Repeat Step 2 until a pre-determined threshold is reached. For final dataset: allocate category by means of a random procedure (repeated 5 times) Thus, 5 final datasets per dimension.

5.6 Conclusion

This concludes the methodology followed for this research. In the following chapter, the results obtained from the simulated data will be presented. The performance of SI and MI with respect to the true data will be illustrated by graphs displaying the relationships between the means and 95% confidence intervals obtained from RIMCA in SI, RIMCA in MI, complete-case analysis and the true data.

Chapter 6
Simulation Study

"Feign,…,pretend to be, act like, resemble, wear the guise of, mimic,…imitate conditions of (situation etc.) with model, for convenience or training…"

(Ripley 1987:1)

6.1 Introduction

This chapter will provide the motivation for using a simulated dataset. The selected dimensions for the reconstruction step of the RIMCA algorithm to generate imputed datasets will be discussed and motivated by means of figures 6.1–6.5. Scatterplot matrices are provided in order to establish the contribution of the initial values. This will be followed by the results obtained from the execution of the two objectives and an overall summary of the performance of the RIMCA MI algorithm over a 1000 simulations; concluding with the discussion of the simulation study results.

6.2 Motivation

The performance and relevance of the RIMCA algorithm in MI will be established when comparing estimates from a complete dataset with the estimates of an imputed version of the original data. The accuracy of the imputation can then be evaluated. According to Ripley (1987:4) the great advantage of simulation is the decrease in approximations, but simultaneously the interpretation of analysis becomes tricky, since the data is not based on true observations and responses.

6.3 Dimensions to retain in the second step of RIMCA

SI

The dimension selected *a priori* for the reconstruction step of the algorithms in SI is based upon the expectance of underfitting (fuzziness) in the lower

dimensions and possible overfitting in the larger dimensions, as mentioned by Josse *et al.* (2012:101). Therefore, an average number of dimensions is selected, which in the case of the simulated data is ten.

MI

As discussed in the preceding chapter (**cf**. 5.5), ten randomly selected dimensions are used to generate the imputed datasets, which are then combined using Rubin's rules. This procedure is repeated ten times. Therefore one repetition refers to the simulation of one dataset, creating missing values within this created dataset according to a specific pattern, percentage of missing values and missingness mechanism. The RIMCA algorithm is then performed on the generated dataset with respect to the randomly selected ten dimensions which will result in ten imputed datasets. These datasets are then combined to determine estimates used in the analysis. The following figures 6.1–6.5 illustrate the means and confidence intervals obtained over the ten repetitions of the sets of randomly selected dimensions for a selection of variables from a dataset with a low correlation structure, MCAR random pattern with a low percentage of missing values. The means of the completed simulated datasets will be displayed and referred to as *Mean Complete*, whereas the mean estimates of the multiply imputed datasets will be indicated by *Q MI*. The figures of the repetitions of all the simulated datasets are presented in Appendix K.

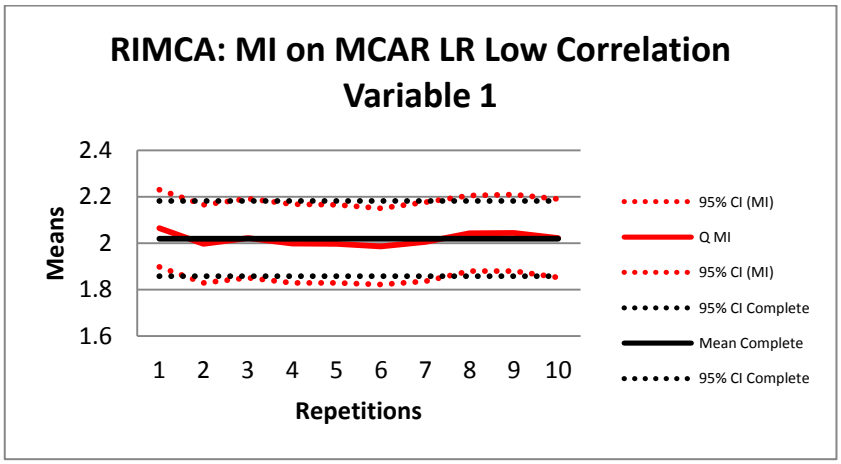


Figure 6.1 RIMCA: MI on MCAR LR data with low correlation structure (variable 1)

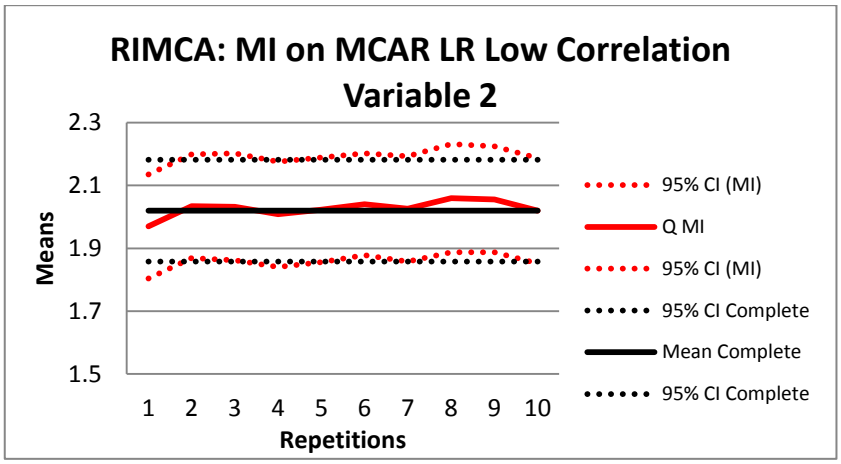


Figure 6.2 RIMCA: MI on MCAR LR data with low correlation structure (variable 2)

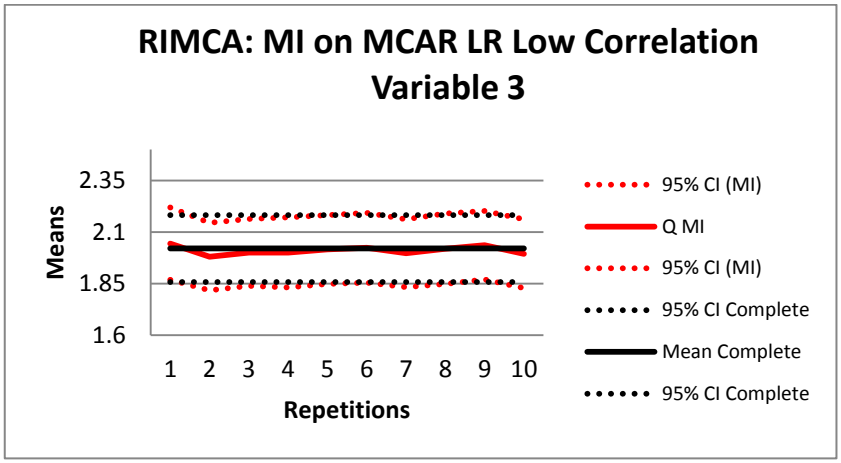


Figure 6.3 RIMCA: MI on MCAR LR data with low correlation structure (variable 3)

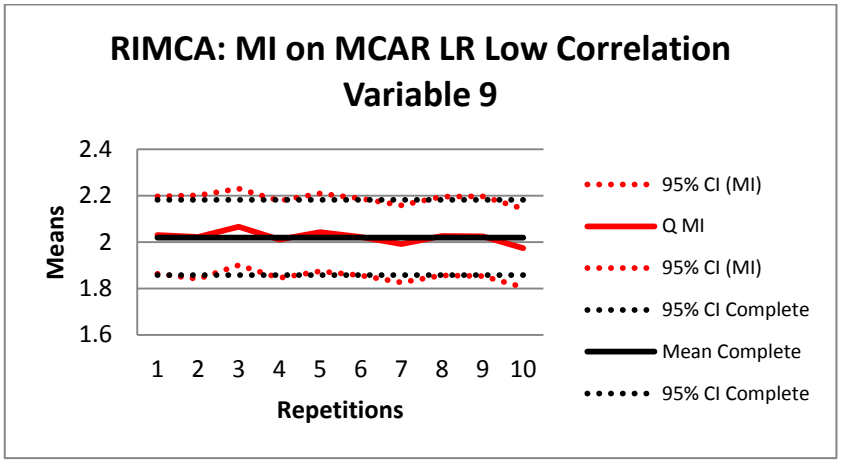


Figure 6.4 RIMCA: MI on MCAR LR data with low correlation structure (variable 9)

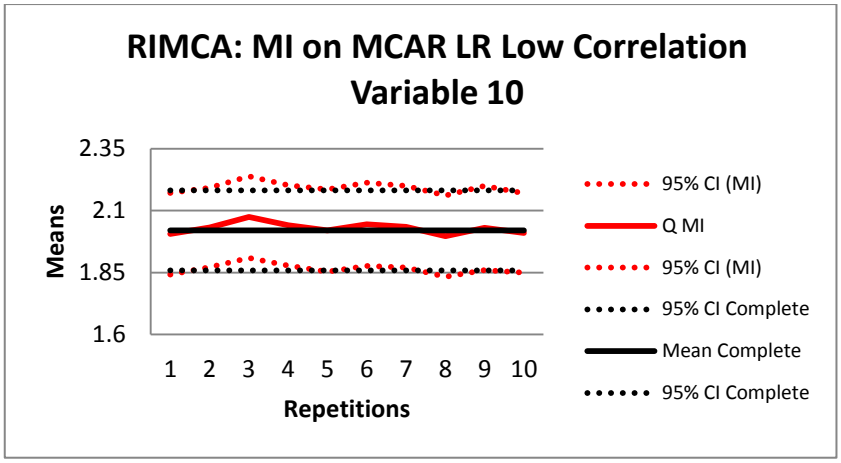


Figure 6.5 RIMCA: MI on MCAR LR data with low correlation structure (variable 10)

It can be seen that the means and confidence intervals stay approximately constant over the ten repetitions; this is the case for all variables and all simulated datasets (**cf**. Appendix K). This confirms that the use of random dimensions for the multiple imputation of the missing values is suitable.

The set of multiple datasets used for further analysis is determined by randomly generated dimensions. The analyses that follow in sections 6.5.1 and 6.5.2 are based on the multiple datasets of one repetition.

6.4 Scatterplot matrices

In order to establish whether the initial fuzzy values allocated in the first step of RIMCA contribute to the final fuzzy values obtained after the convergence of the reconstruction steps, scatterplot matrices are used to display the relationship between the multiple datasets with regard to the fuzzy values obtained for categories one and two.

The scatterplot matrices of the MCAR missingness mechanism with a low percentage of missing values and a random pattern are displayed for the high correlation structure (**cf**. figure 6.6) and the low correlation structure (**cf**. figure 6.7).

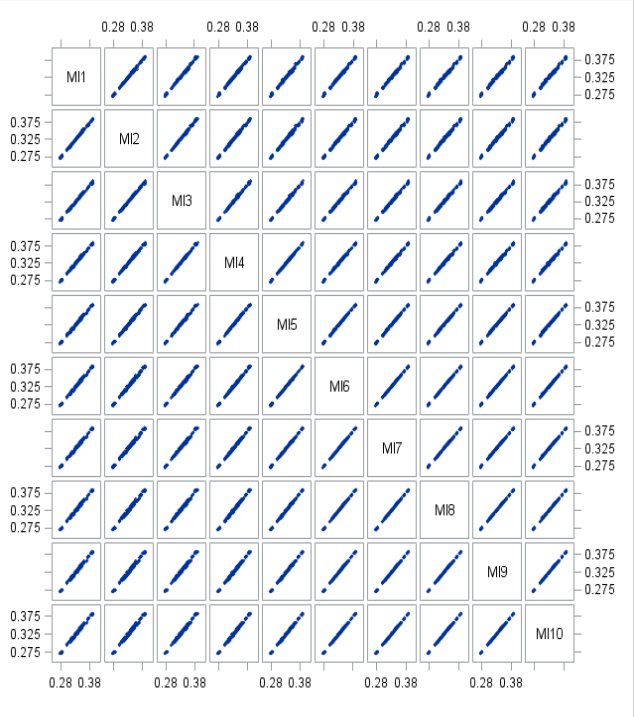


Figure 6.6 Scatterplot matrix of MCAR LR data with a high correlation structure

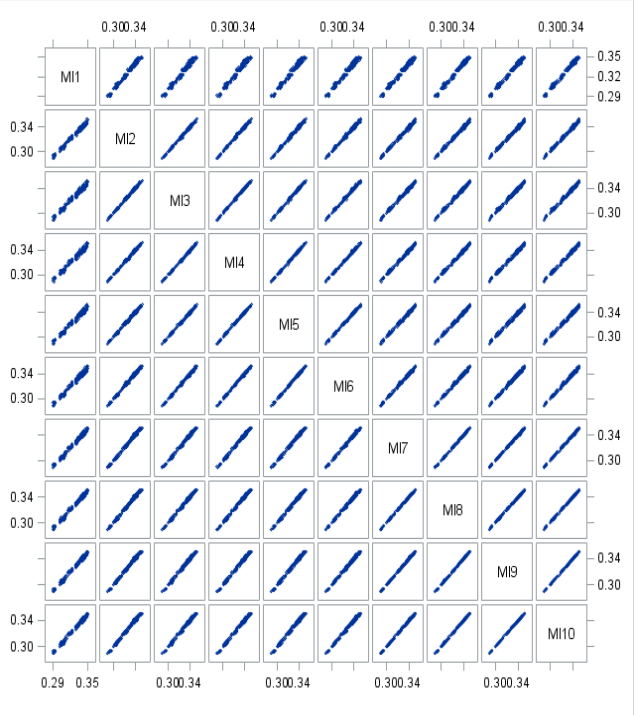


Figure 6.7 Scatterplot matrix of MCAR LR data with a low correlation structure

The scatterplots show a 45 degree line for all matrices, which confirms the statement of Josse *et al.* (2012:100) that the initial values do not contribute in any way and will therefore not influence the imputations. This also confirms that the RIMCA algorithm is not sensitive to random starting values, as was

predicted by Josse *et al.* (2012:100). The scatterplots show that the RIMCA algorithm provides the same or similar fuzzy values for each observation across the ten multiple datasets, obtained from a set of ten randomly selected dimensions. This shows that the final datasets will always be similar, irrespective of the chosen dimension. Scatterplots of the other simulated datasets are available in Appendix N.

6.5 Objective one: To establish whether RIMCA in MI outperforms RIMCA in SI

In the figures that will follow in sections 6.5.1 and 6.5.2 the means of the completed simulated datasets will be displayed and referred to as *CD Mean* or the true mean and the *CD Confidence Intervals* will be referred to as the true confidence intervals. The estimates of the mean obtained from complete-case analysis will be referred to as *CC Mean*. The estimated means obtained from the imputation processes will be indicated by *MI Mean* and *SI Mean*, indicating the MI estimate and the SI estimate, respectively. As was discussed in the simulation protocol (**cf**. 5.4.1.1), the high correlation structure refers to a correlation of 0.8 and a low correlation of 0.4. The missing values are entered using a random (R) or non-random (NR) pattern. The first section will display the means and confidence intervals of the datasets with MAR values, followed by the section on the datasets with MCAR values. Each sub-section will consist of a table and three figures for both correlation structures and each particular pattern (random or non-random). The information obtained for one variable from the application of Rubin's rules (**cf**. 3.5) for the MI case and the results obtained for the SI procedure will be presented with the true values from the completed data. The first figure represents the means and confidence intervals of the completed data, MI, SI and complete-case (CC) procedures on the same graph. The second and third graphs will show the performance of MI and SI, respectively, in comparison to the mean and confidence intervals obtained from the completed data and the complete-case analysis.

6.5.1 Simulated data with a MAR missingness mechanism

The datasets with a MAR mechanism either have a low percentage (L) of missing values (8%) or a high percentage (H) of missing values (16%). In all of the MAR instances variable one and seven are completely observed, therefore no estimates are obtained for these variables. The confidence intervals for these variables are equal to the true confidence intervals of the original simulated data in both the SI and MI cases. The analysis will be restricted to imputed variables only.

6.5.1.1 MAR HR High correlation structure

Table 6.1 Confidence interval widths, means and standard errors obtained from complete-case analysis and RIMCA in SI and MI for MAR HR high correlated data in comparison to the true values

MAR HR High	Confidence Interval Width				Mean			
Variable	CD	CC	SI	MI	CD	CC	SI	MI
1	0.3239	observed	observed	observed	2.02	observed	observed	observed
2*	0.3239	0.3424	0.2973	**0.3208**	2.02	2.2208	2.3800	2.2150
3*	0.3239	0.3611	0.3083	**0.3445**	2.02	2.1948	2.3200	2.1908
4*	0.3239	0.3424	0.3019	**0.3294**	2.02	2.2208	2.3700	2.2112
5*	0.3239	0.3611	0.3012	**0.3378**	2.02	2.1948	2.3600	2.1998
6*	0.3239	0.3407	0.2649	**0.3315**	2.02	2.2597	2.1700	2.2650
7	0.3239	observed	observed	observed	2.02	observed	observed	observed
8*	0.3239	0.3557	0.3049	**0.3449**	2.02	1.9157	1.6600	1.9196
9*	0.3239	0.3530	0.3134	**0.3465**	2.02	1.9277	1.6800	1.9400
10*	0.3239	0.3622	0.3092	**0.3510**	2.02	1.9157	1.6700	1.9202

MAR HR High	Standard Errors			
Variable	CD	CC	SI	MI
1	0.0816	0.0816	0.0816	0.0816
2	0.0816	0.0860	0.0749	0.0818
3	0.0816	0.0906	0.0777	0.0878
4	0.0816	0.0860	0.0761	0.0840
5	0.0816	0.0906	0.0759	0.0861
6	0.0816	0.0855	0.0667	0.0845
7	0.0816	0.0816	0.0816	0.0816
8	0.0816	0.0894	0.0768	0.0879
9	0.0816	0.0887	0.0790	0.0883
10	0.0816	0.0910	0.0779	0.0895

CD – complete data, CC – complete-case analysis, SI – single imputation, MI – multiple imputation
* – indicates the variables with a wider confidence interval with regard to MI (only considering SI and MI)

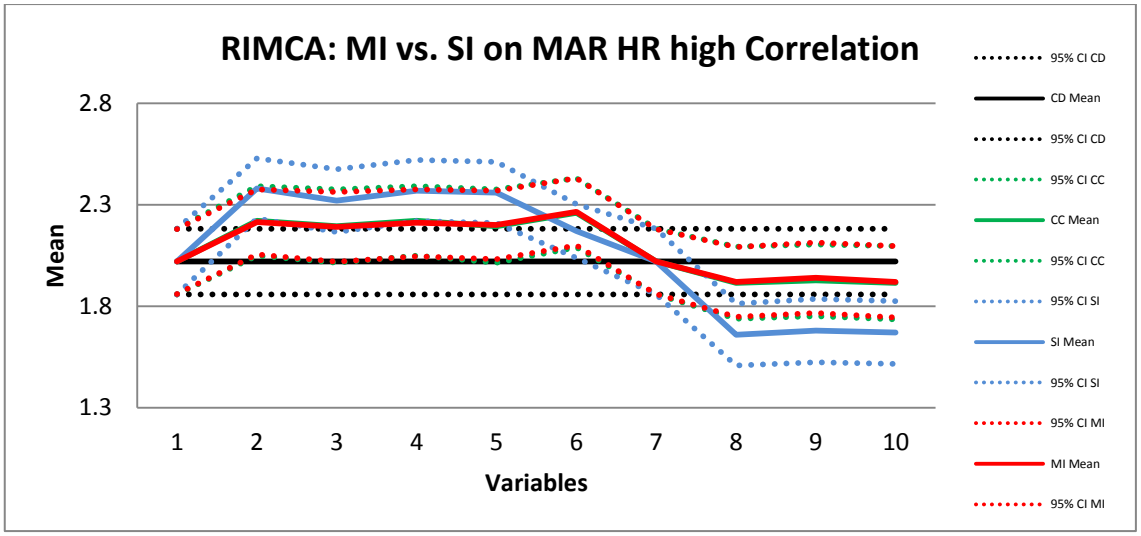


Figure 6.8 Means and Confidence intervals for RIMCA in MI and SI (MAR HR)

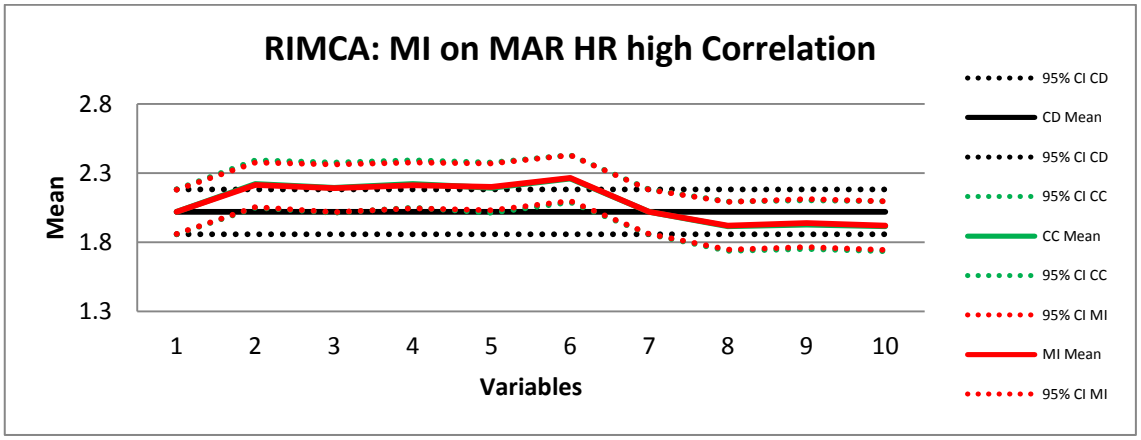


Figure 6.9 MI and CC vs. CD Mean and CI's on MAR HR High correlated data

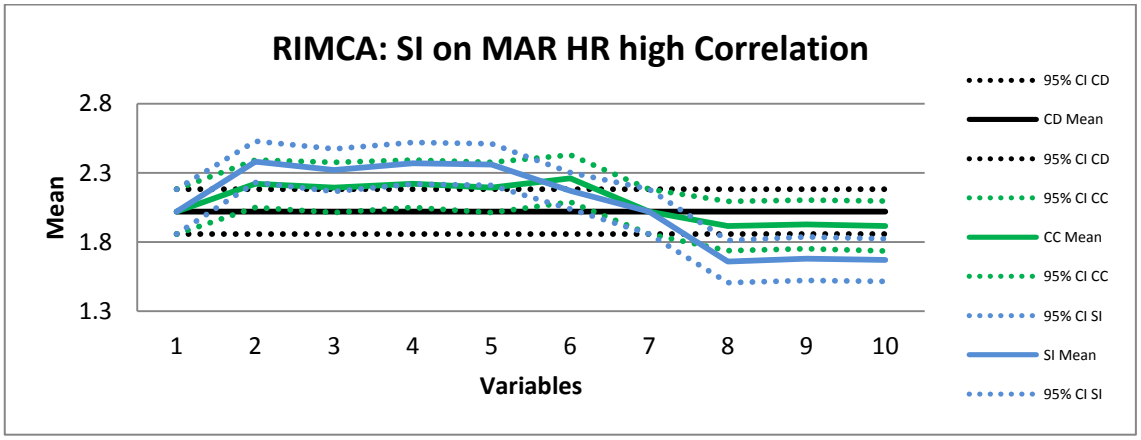


Figure 6.10 SI and CC vs. CD Mean and CI's on MAR HR High correlated data

Discussion

Table 6.1 indicates that MI produces wider confidence intervals than SI for all variables. The graphical representation of this information is provided in figure 6.8. Figures 6.8 and 6.10 show that there is strong evidence that SI is statistically different from the true confidence intervals for variables two, four, five, eight, nine and ten, since the intervals for these variables do not overlap the true confidence intervals. Therefore, for these specified variables SI will not provide accurate predictions. The mean estimate obtained from SI for variable six is closer to the true mean than the mean estimated obtained from MI, whereas all other mean estimates provided by MI (**cf**. figure 6.9) are closer to the true mean than SI. The CC estimates are closely correlated to the MI estimates (**cf**. figure 6.9) and for all of the imputed variables wider confidence intervals are obtained from the CC analysis in comparison to the MI procedure. Larger standard errors are obtained from the CC analysis for all of the variables.

Thus it is clear that MI is a better fit for the MAR high correlated data with a random pattern and higher percentage of missing values than SI.

6.5.1.2 MAR HR Low correlation structure

Table 6.2 Confidence interval widths, means and standard errors obtained from complete-case analysis and RIMCA in SI and MI for MAR HR low correlated data in comparison to the true values

MAR HR Low	Confidence Interval Width				Mean			
Variable	CD	CC	SI	MI	CD	CC	SI	MI
1	0.3239	observed	observed	observed	2.02	observed	observed	observed
2*	0.3239	0.3565	0.3282	**0.3403**	2.02	2.0625	2.2700	2.0652
3*	0.3239	0.3611	0.3295	**0.3428**	2.02	2.1125	2.2400	2.1120
4*	0.3239	0.3713	0.3195	**0.3595**	2.02	2.0125	2.2800	2.0198
5*	0.3239	0.3678	0.3233	**0.3527**	2.02	2.0250	2.2700	2.0352
6*	0.3239	0.3645	0.2848	**0.3412**	2.02	2.0125	2.1000	2.0110
7	0.3239	observed	observed	observed	2.02	observed	observed	observed
8*	0.3239	0.3635	0.2816	**0.3427**	2.02	1.9375	1.9600	1.9412
9*	0.3239	0.3565	0.3191	**0.3464**	2.02	1.9375	2	1.9440
10*	0.3239	0.3534	0.3134	**0.3397**	2.02	1.9500	1.6800	1.9616

MAR HR Low	Standard Errors			
Variable	CD	CC	SI	MI
1	0.0816	0.0816	0.0816	0.0816
2	0.0816	0.0896	0.0827	0.0868
3	0.0816	0.0907	0.0830	0.0874
4	0.0816	0.0933	0.0805	0.0916
5	0.0816	0.0924	0.0815	0.0899
6	0.0816	0.0916	0.0718	0.0870
7	0.0816	0.0816	0.0816	0.0816
8	0.0816	0.0913	0.0710	0.0874
9	0.0816	0.0896	0.0804	0.0883
10	0.0816	0.0888	0.0790	0.0866

CD – complete data, CC – complete-case analysis, SI – single imputation, MI – multiple imputation
* – indicates the variables with a wider confidence interval with regard to MI (only considering SI and MI)

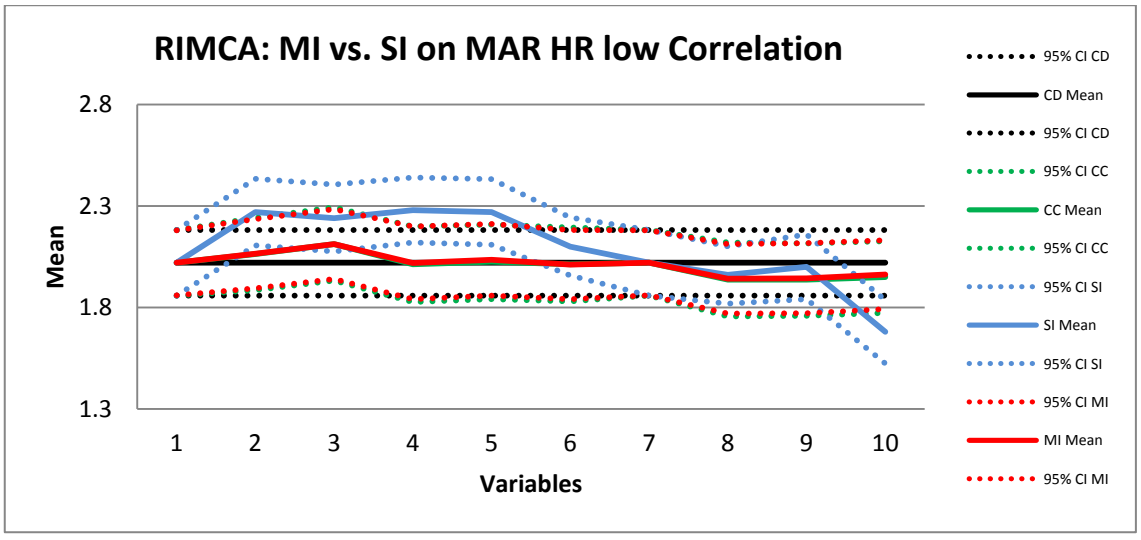


Figure 6.11 Means and Confidence intervals for RIMCA in MI and SI (MAR HR)

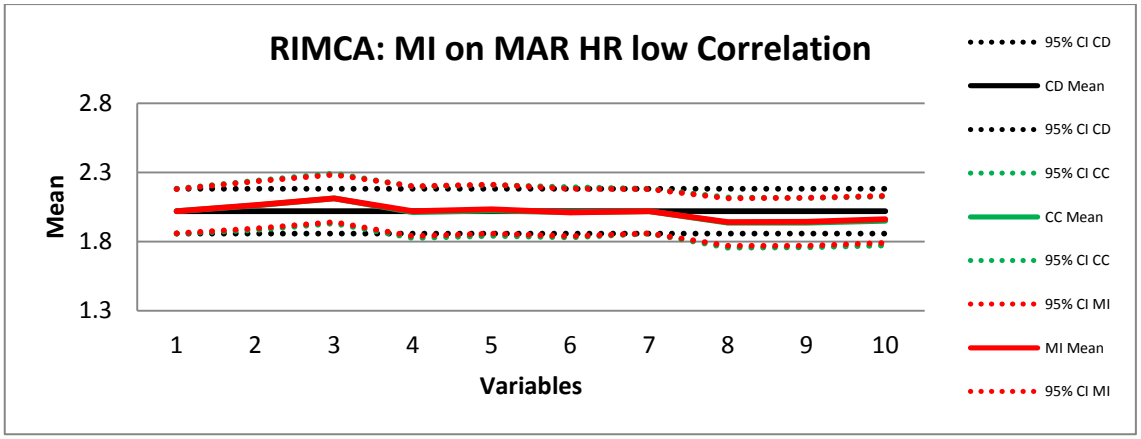


Figure 6.12 MI and CC vs. CD Mean and CI's on MAR HR Low correlated data

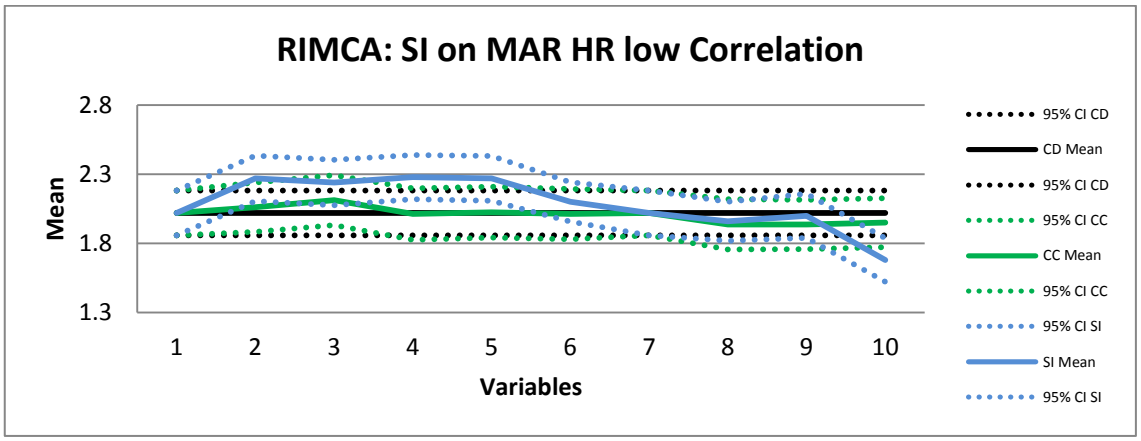


Figure 6.13 SI and CC vs. CD Mean and CI's on MAR HR Low correlated data

Discussion

Table 6.2 indicates that MI produces wider confidence intervals than SI for all variables. The graphical representation of this information is provided in figure 6.11. The mean estimates obtained from SI (**cf**. figures 6.11 & 6.13) for variables six and nine are closer to the true mean than the MI estimates for these variables, whereas all other mean estimates provided by MI (**cf**. figures 6.11 & 6.12) are closer to the true mean than SI. As was the case with the high correlation structure (**cf**. 6.5.1.1), the MI and CC estimates are closely correlated (**cf**. figure 6.11). The confidence intervals widths are larger for the CC estimates than the MI estimates, as are standard errors.

Thus it is clear that MI is a better fit for the MAR low correlated data with a random pattern and high percentage of missing values than SI.

Conclusion for MAR HR for High- and Low correlated data

In both high and low correlated data RIMCA in MI is a better fit than RIMCA in SI. This is confirmed by the statistically different confidence intervals obtained from SI for some of the variables in the high correlated data. Both the MI and SI procedures perform better under the low correlation structure. A great advantage is the fact that wider confidence intervals are obtained from MI in both correlation structures, which incorporates the uncertainty when imputing missing values. The estimates obtained from CC are similar to the MI estimates and perform better than the SI procedure. Therefore, RIMCA in MI is a better imputation model than RIMCA in SI for the specified data.

6.5.1.3 MAR HNR High correlation structure

Table 6.3 Confidence interval widths, means and standard errors obtained from complete-case analysis and RIMCA in SI and MI for MAR HNR high correlated data in comparison to the true values

MAR HNR High	Confidence Interval Width				Mean			
Variable	CD	CC	SI	MI	CD	CC	SI	MI
1	0.3239	observed	observed	observed	2.02	observed	observed	observed
2*	0.3239	0.3453	0.3057	**0.3296**	2.02	2.1750	2.3500	2.1794
3*	0.3239	0.3525	0.2743	**0.3561**	2.02	2.1750	2.1300	2.1782
4*	0.3239	0.3320	0.2973	**0.3291**	2.02	2.2250	2.3800	2.2182
5*	0.3239	0.3426	0.3012	**0.3306**	2.02	2.2000	2.3600	2.2036
6*	0.3239	0.3284	0.2796	**0.3145**	2.02	2.2500	2.2200	2.2512
7	0.3239	observed	observed	observed	2.02	observed	observed	observed
8*	0.3239	0.3523	0.3049	**0.3370**	2.02	1.8625	1.6600	1.8542
9*	0.3239	0.3451	0.3092	**0.3274**	2.02	1.8625	1.6700	1.8578
10*	0.3239	0.3593	0.3049	**0.3386**	2.02	1.8625	1.6600	1.8656

MAR HNR High	Standard Errors			
Variable	CD	CC	SI	MI
1	0.0816	0.0816	0.0816	0.0816
2	0.0816	0.0867	0.0770	0.0840
3	0.0816	0.0885	0.0691	0.0907
4	0.0816	0.0834	0.0749	0.0839
5	0.0816	0.0861	0.0759	0.0843
6	0.0816	0.0825	0.0705	0.0802
7	0.0816	0.0816	0.0816	0.0816
8	0.0816	0.0885	0.0768	0.0859
9	0.0816	0.0867	0.0779	0.0835
10	0.0816	0.0903	0.0768	0.0863

CD – complete data, CC – complete-case analysis, SI – single imputation, MI – multiple imputation
* – indicates the variables with a wider confidence interval with regard to MI (only considering SI and MI)

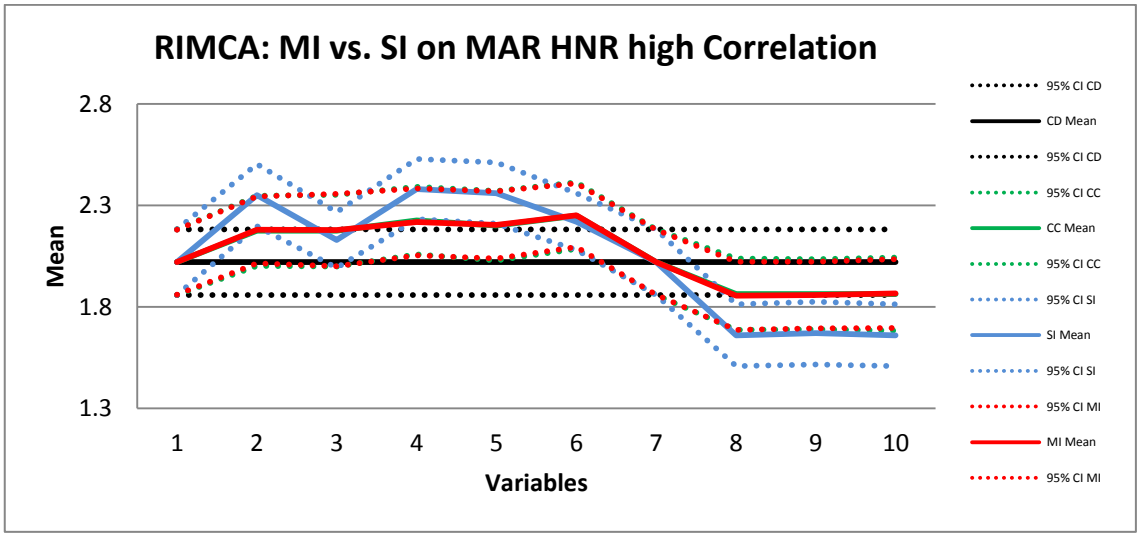


Figure 6.14 Means and Confidence intervals for RIMCA in MI and SI (MAR HNR)

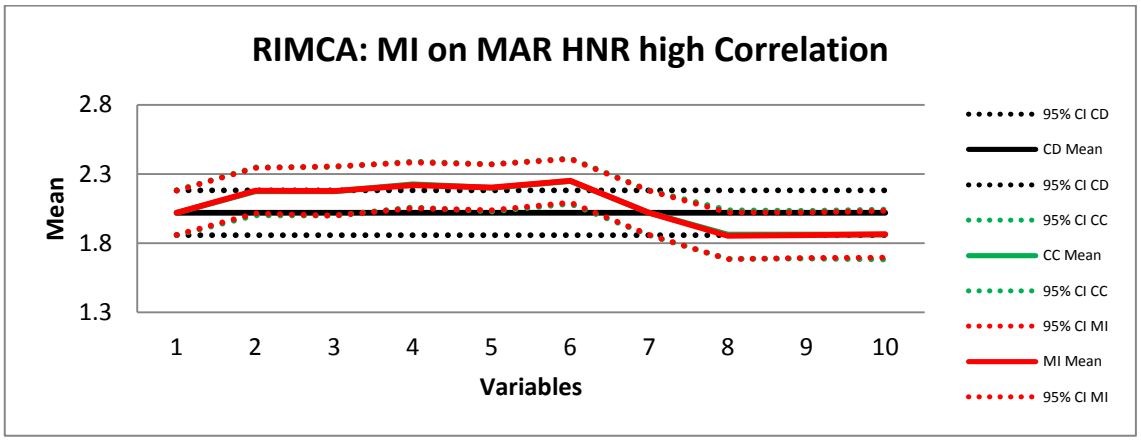


Figure 6.15 MI and CC vs. CD Mean and CI's on MAR HNR High correlated data

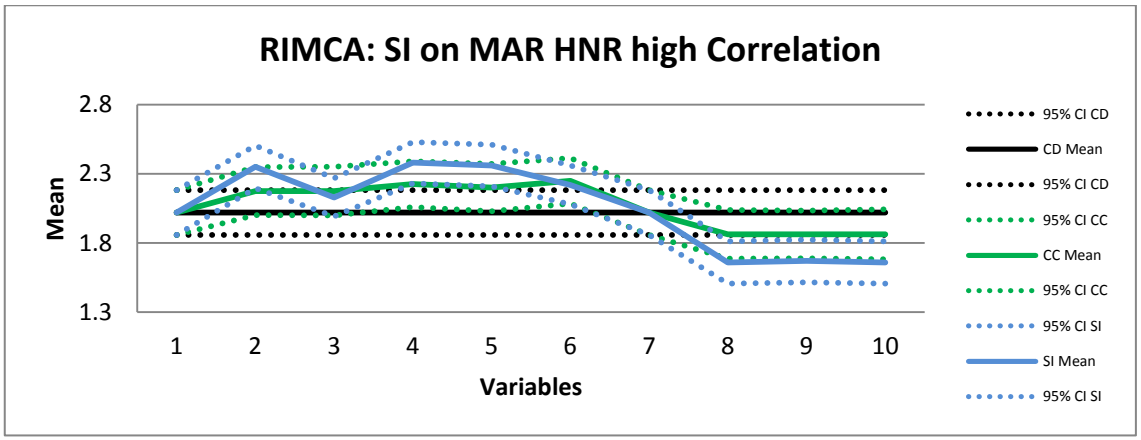


Figure 6.16 SI and CC vs. CD Mean and CI's on MAR HNR High correlated data

Discussion

Table 6.3 indicates that MI produces wider confidence intervals than SI for all variables. CC produces wider confidence intervals than MI for all of the variables, except variable three. Also the standard error obtained from MI for variable three is larger than the CC standard error for this specific variable. Overall the CC analysis provides larger standard errors than the imputation techniques. The graphical representation of this information is provided in figure 6.14. Figures 6.14 and 6.16 show that there is strong evidence that SI is statistically different from the true confidence intervals for variables four, five, eight, nine and ten, since the intervals for these variables do not overlap the true intervals. Therefore, for these specified variables SI will not provide accurate predictions. The mean estimate obtained from SI for variable three is slightly closer to the true mean than the MI estimate for this variable. However, the MI estimates (**cf**. figures 6.14 & 6.15) obtained for all the remaining variables are closer to the true mean value than SI. The MI estimates and confidence intervals are similar to the CC estimates (**cf**. figure 6.15).

The estimates obtained from MI perform better than the SI estimates and MI seems to be a good fit for the prediction of the missing values in this particular dataset.

6.5.1.4 MAR HNR Low correlation structure

Table 6.4 Confidence interval widths, means and standard errors obtained from complete-case analysis and RIMCA in SI and MI for MAR HNR low correlated data in comparison to the true values

MAR HNR Low	Confidence Interval Width				Mean			
Variable	CD	CC	SI	MI	CD	CC	SI	MI
1	0.3239	observed	observed	observed	2.02	observed	observed	observed
2*	0.3239	0.3664	0.3282	**0.3488**	2.02	2.0750	2.2700	2.0746
3*	0.3239	0.3637	0.3205	**0.3487**	2.02	2.1250	2.2900	2.1304
4*	0.3239	0.3741	0.3259	**0.3674**	2.02	2.0500	2.2500	2.0582
5*	0.3239	0.3565	0.3205	**0.3472**	2.02	2.0625	2.2900	2.0602
6*	0.3239	0.3642	0.3271	**0.3594**	2.02	2.0375	2.2600	2.0358
7	0.3239	observed	observed	observed	2.02	observed	observed	observed
8*	0.3239	0.3642	0.3109	**0.3464**	2.02	1.9625	1.8500	1.9740
9*	0.3239	0.3555	0.3282	**0.3413**	2.02	1.9125	1.7700	1.9132
10*	0.3239	0.3555	0.3195	**0.3474**	2.02	1.9125	1.7200	1.9120

MAR HNR Low	Standard Errors			
Variable	CD	CC	SI	MI
1	0.0816	0.0816	0.0816	0.0816
2	0.0816	0.0920	0.0827	0.0889
3	0.0816	0.0914	0.0808	0.0889
4	0.0816	0.0940	0.0821	0.0936
5	0.0816	0.0896	0.0808	0.0885
6	0.0816	0.0915	0.0824	0.0916
7	0.0816	0.0816	0.0816	0.0816
8	0.0816	0.0915	0.0783	0.0883
9	0.0816	0.0893	0.0827	0.0870
10	0.0816	0.0893	0.0805	0.0886

CD – complete data, CC – complete-case analysis, SI – single imputation, MI – multiple imputation
* – indicates the variables with a wider confidence interval with regard to MI (only considering SI and MI)

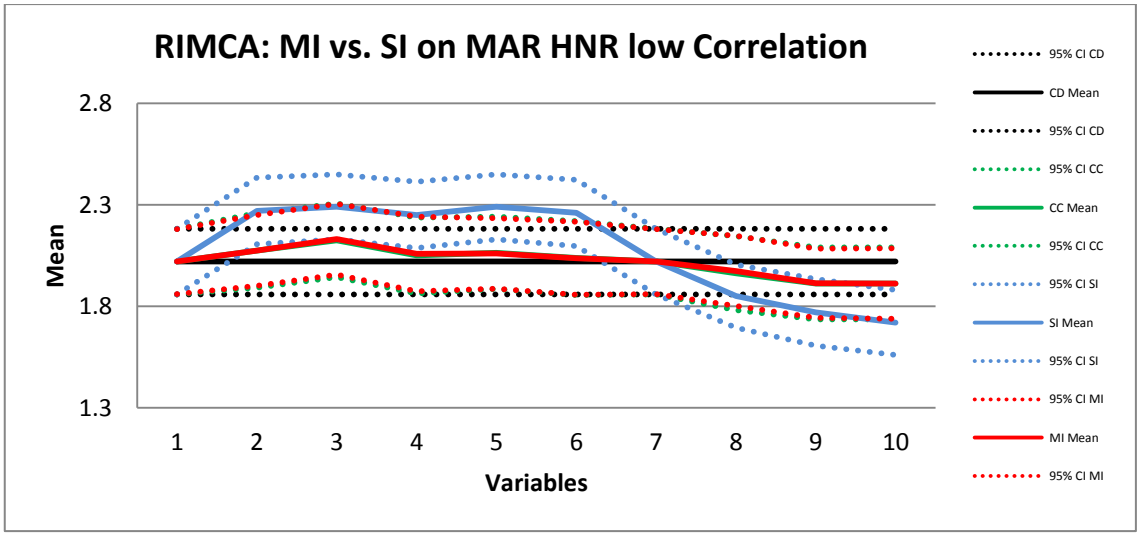


Figure 6.17 Means and Confidence intervals for RIMCA in MI and SI (MAR HNR)

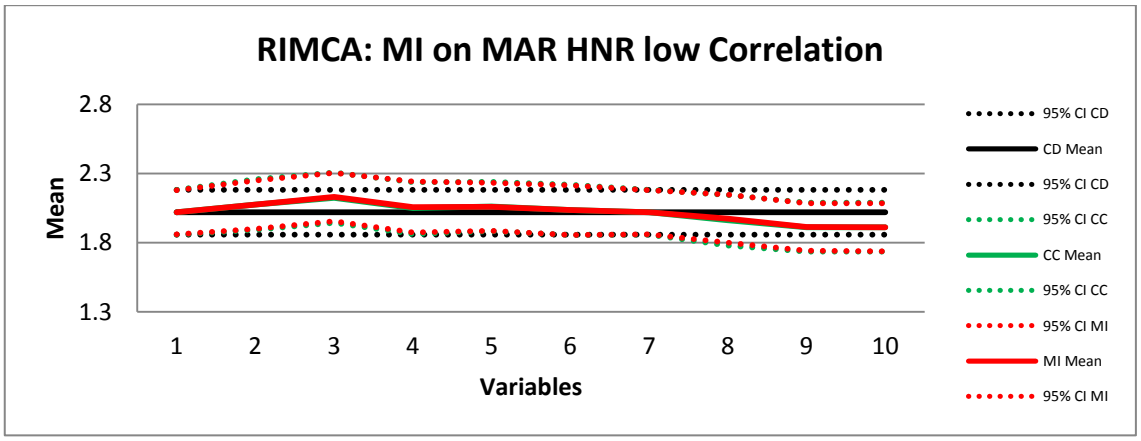


Figure 6.18 MI and CC vs. CD Mean and CI's on MAR HNR Low correlated data

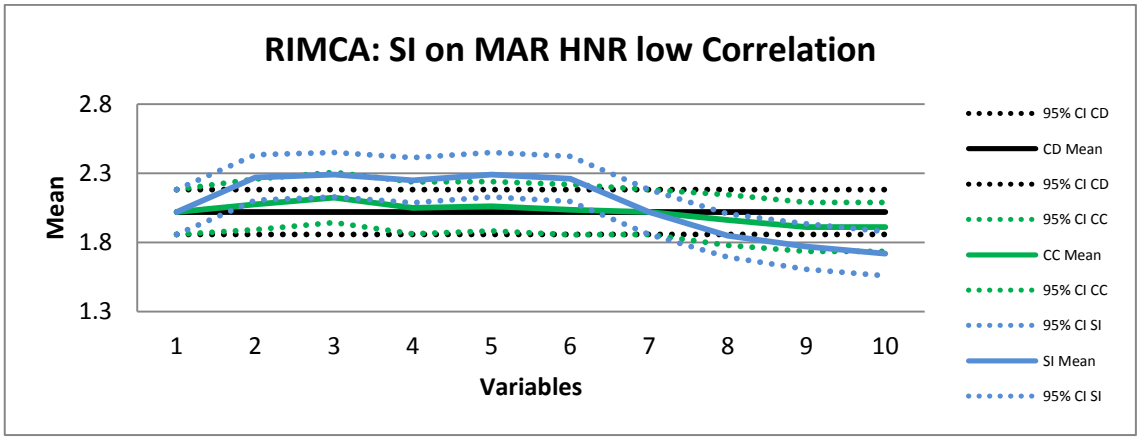


Figure 6.19 SI and CC vs. CD Mean and CI's on MAR HNR Low correlated data

Discussion

Table 6.4 indicates that MI produces wider confidence intervals than SI for all variables. The graphical representation of this information is provided in figure 6.17. The mean estimates obtained from MI for all variables are closer to the true mean than the estimates obtained from SI. The SI predictions (**cf**. figure 6.19) are more unstable between variables than the mean values predicted by MI with respect to the true values (**cf**. figure 6.18). The CC results are closely correlated with the MI results.

MI is a better fit for the MAR low correlated data with a non-random pattern and high percentage of missing values than SI.

Conclusion for MAR HNR for High- and Low correlated data

RIMCA in MI performed better than RIMCA in SI, since the mean estimates resembles the true means more closely. The wider confidence intervals obtained from MI in both correlation structures confirm the added uncertainty when imputing with multiple datasets. SI performs better in the low correlated data, since the confidence intervals for all the variables overlap with the true confidence intervals. MI provides better estimates than SI in all cases and is therefore a better fit for imputation of this specific dataset. RIMCA in MI provided slightly better results in the presence of a low correlation structure. RIMCA in MI is a better imputation model than RIMCA in SI for the specified data.

6.5.1.5 MAR LR High correlation structure

Table 6.5 Confidence interval widths, means and standard errors obtained from complete-case analysis and RIMCA in SI and MI for MAR LR high correlated data in comparison to the true values

MAR LR High	Confidence Interval Width				Mean			
Variable	CD	CC	SI	MI	CD	CC	SI	MI
1	0.3239	observed	observed	observed	2.02	observed	observed	observed
2*	0.3239	0.3403	0.3155	**0.3306**	2.02	2.0556	2.2100	2.0524
3*	0.3239	0.3326	0.3225	**0.3278**	2.02	2.1000	2.1900	2.0920
4*	0.3239	0.3361	0.2946	**0.3262**	2.02	2.0889	2.1200	2.0946
5*	0.3239	0.3361	0.3225	**0.3309**	2.02	2.0889	2.1900	2.0810
6*	0.3239	0.3349	0.3105	**0.3281**	2.02	2.1111	2.2100	2.1108
7	0.3239	observed	observed	observed	2.02	observed	observed	observed
8*	0.3239	0.3339	0.2998	**0.3231**	2.02	1.9101	1.9300	1.9164
9*	0.3239	0.3374	0.2998	**0.3246**	2.02	1.9213	1.9300	1.9256
10*	0.3239	0.3408	0.3290	**0.3293**	2.02	1.9326	1.8600	1.9260

MAR LR High	Standard Errors			
Variable	CD	CC	SI	MI
1	0.0816	0.0816	0.0816	0.0816
2	0.0816	0.0856	0.0795	0.0843
3	0.0816	0.0837	0.0813	0.0836
4	0.0816	0.0846	0.0742	0.0832
5	0.0816	0.0846	0.0813	0.0844
6	0.0816	0.0843	0.0782	0.0837
7	0.0816	0.0816	0.0816	0.0816
8	0.0816	0.0840	0.0756	0.0824
9	0.0816	0.0849	0.0756	0.0828
10	0.0816	0.0858	0.0829	0.0840

CD – complete data, CC – complete-case analysis, SI – single imputation, MI – multiple imputation
* – indicates the variables with a wider confidence interval with regard to MI (only considering SI and MI)

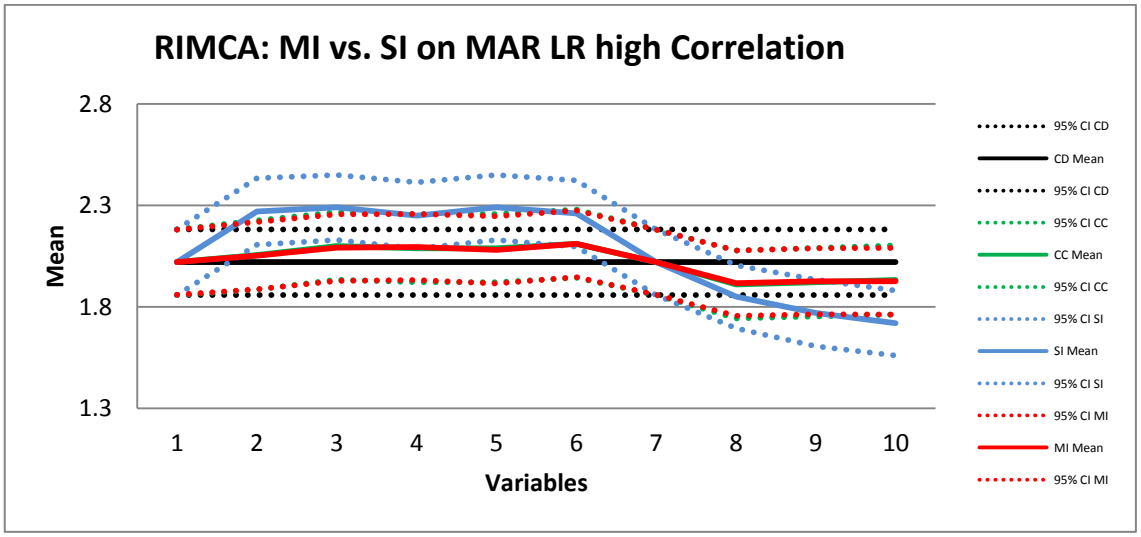


Figure 6.20 Means and Confidence intervals for RIMCA in MI and SI (MAR LR)

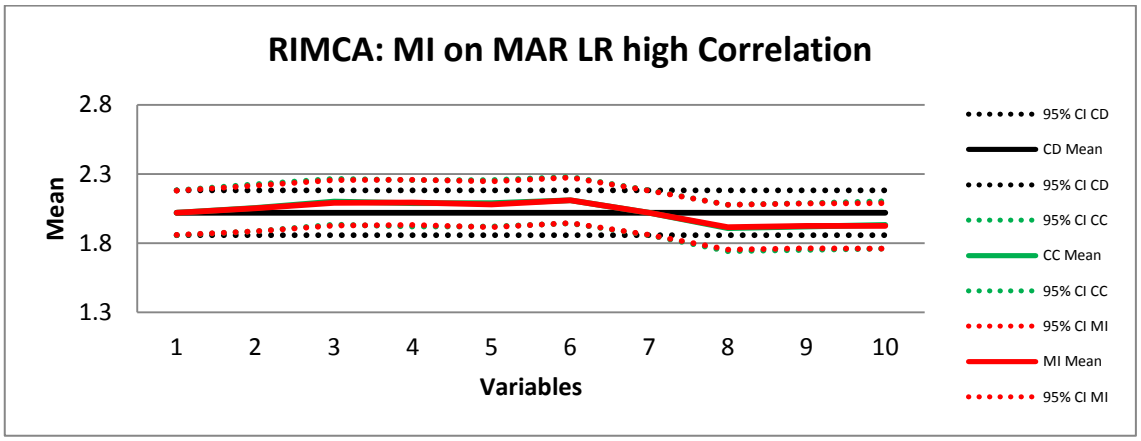


Figure 6.21 MI and CC vs. CD Mean and CI's on MAR LR High correlated data

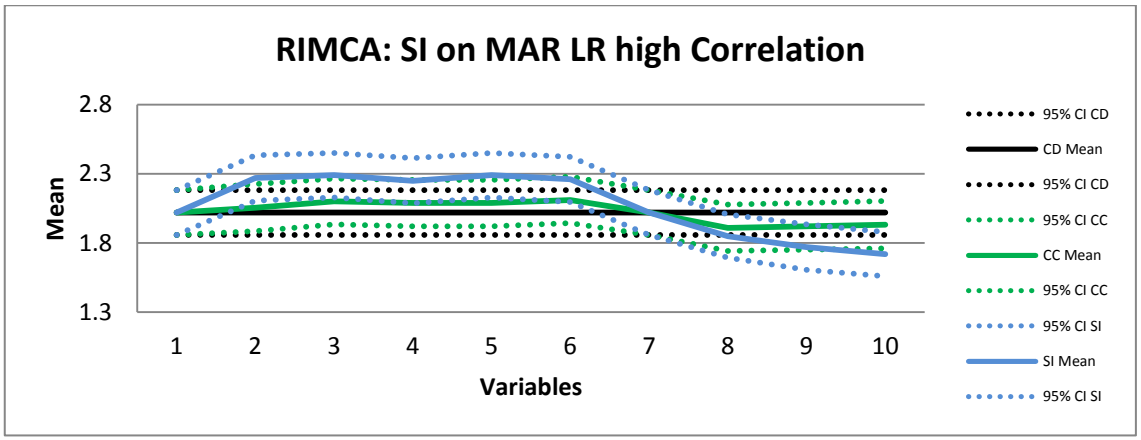


Figure 6.22 SI and CC vs. CD Mean and CI's on MAR LR High correlated data

Discussion

Table 6.5 indicates that MI produces wider confidence intervals than SI for all variables. Further it can be observed that CC provides even wider confidence intervals than MI for all variables. The graphical representation of this information is provided in figure 6.20. The mean estimates obtained from SI for variable four and nine are slightly closer to the true mean than the MI estimates. However, MI provides better estimates for all of the remaining variables. The MI estimates and CC estimates are similar for all variables (**cf**. figure 6.21). The SI predictions (**cf**. figure 6.22) are more unstable between variables than the mean values predicted by MI (**cf**. figure 6.21).

Thus it is clear that MI is a better fit for the MAR high correlated data with a random pattern and low percentage of missing values than SI.

6.5.1.6 MAR LR Low correlation structure

Table 6.6 Confidence interval widths, means and standard errors obtained from complete-case analysis and RIMCA in SI and MI for MAR LR low correlated data in comparison to the true values

MAR LR Low	Confidence Interval Width				Mean			
Variable	CD	CC	SI	MI	CD	CC	SI	MI
1	0.3239	observed	observed	observed	2.02	observed	observed	observed
2*	0.3239	0.3382	0.3243	**0.3271**	2.02	2	2.1700	1.9978
3*	0.3239	0.3495	0.3274	**0.3421**	2.02	2.0222	2.1900	2.0198
4*	0.3239	0.3439	0.3243	**0.3352**	2.02	2	2.1700	2.0006
5*	0.3239	0.3403	0.3225	**0.3353**	2.02	2.0556	2.1900	2.0538
6*	0.3239	0.3438	0.3225	**0.3377**	2.02	2.0222	2.1900	2.0164
7	0.3239	observed	observed	observed	2.02	observed	observed	observed
8*	0.3239	0.3478	0.3061	**0.3384**	2.02	2.0225	1.9700	2.0270
9*	0.3239	0.3449	0.3302	**0.3347**	2.02	1.9888	1.8800	1.9924
10*	0.3239	0.3382	0.3272	**0.3306**	2.02	1.9438	1.8700	1.9402

MAR LR Low	Standard Errors			
Variable	CD	CC	SI	MI
1	0.0816	0.0816	0.0816	0.0816
2	0.0816	0.0851	0.0817	0.0834
3	0.0816	0.0879	0.0825	0.0873
4	0.0816	0.0865	0.0817	0.0855
5	0.0816	0.0856	0.0813	0.0855
6	0.0816	0.0865	0.0813	0.0861
7	0.0816	0.0816	0.0816	0.0816
8	0.0816	0.0875	0.0771	0.0863
9	0.0816	0.0868	0.0832	0.0854
10	0.0816	0.0851	0.0825	0.0843

CD – complete data, CC – complete-case analysis, SI – single imputation, MI – multiple imputation
* – indicates the variables with a wider confidence interval with regard to MI (only considering SI and MI)

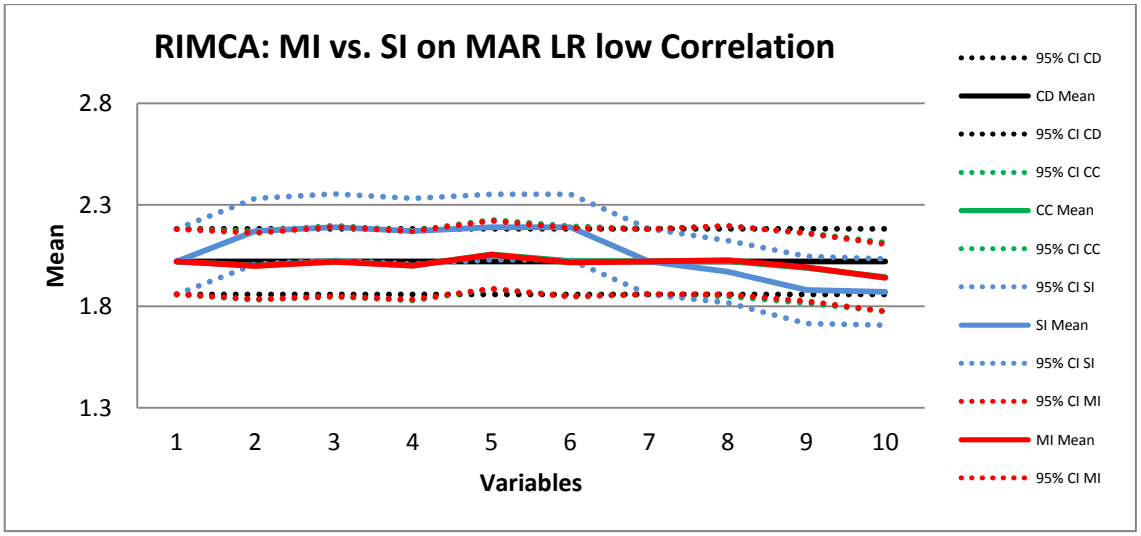


Figure 6.23 Means and Confidence intervals for RIMCA in MI and SI (MAR LR)

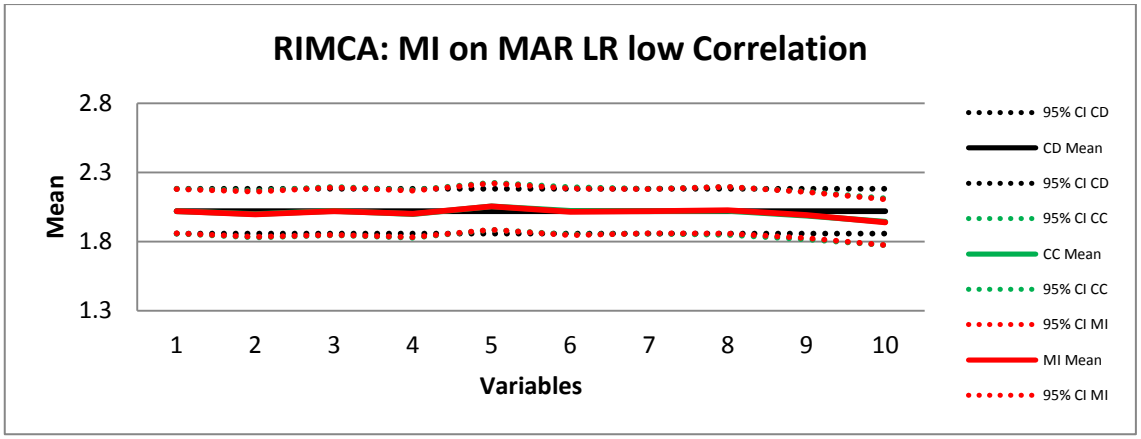


Figure 6.24 MI and CC vs. CD Mean and CI's on MAR LR Low correlated data

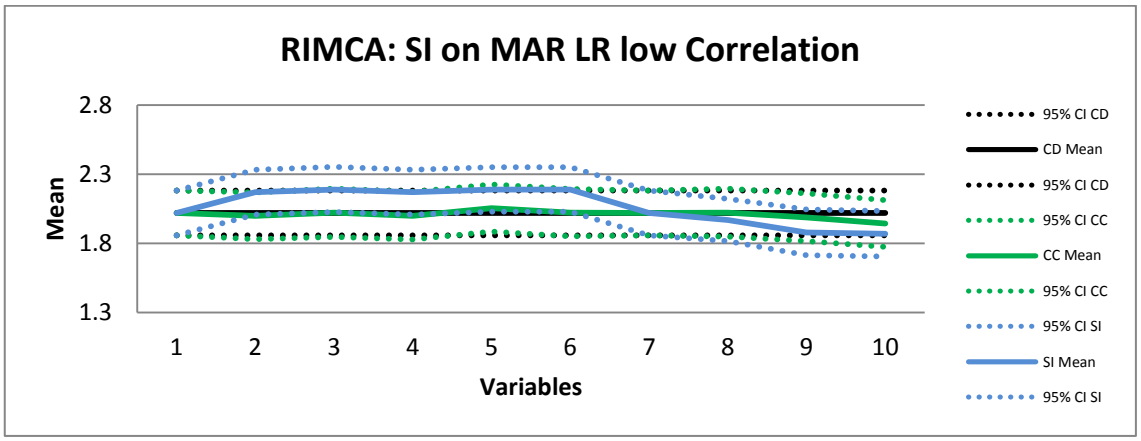


Figure 6.25 SI and CC vs. CD Mean and CI's on MAR LR Low correlated data

Discussion

Table 6.6 indicates that MI produces wider confidence intervals than SI for all variables. The graphical representation of this information is provided in figure 6.23. The results obtained from CC and MI are similar with respect to the mean estimates, but larger standard errors and wider confidence intervals are obtained from CC. The mean estimates obtained from MI in comparison to the estimates obtained from SI, are closer to the true mean values in all of the variables (**cf**. figures 6.23 & 6.24).

Thus it is clear that MI is a better fit for the MAR low correlated data with a non-random pattern and low percentage of missing values than SI (**cf**. figures 6.23 & 6.25).

Conclusion for MAR LR for High- and Low correlated data

SI seems to perform slightly better in the presence of a high correlation structure in the data, whereas MI performs better for low correlated data. MI outperforms SI with regard to wider confidence intervals and estimates that are closer to the true mean values.

RIMCA in MI is a better imputation model than RIMCA in SI for the specified data.

6.5.1.7 MAR LNR High correlation structure

Table 6.7 Confidence interval widths, means and standard errors obtained from complete-case analysis and RIMCA in SI and MI for MAR LNR high correlated data in comparison to the true values

MAR LNR High	**Confidence Interval Width**				**Mean**			
Variable	CD	CC	SI	MI	CD	CC	SI	MI
1	0.3239	observed	observed	observed	2.02	observed	observed	observed
2*	0.3239	0.3326	0.3210	**0.3252**	2.02	2.1000	2.1800	2.0984
3*	0.3239	0.3419	0.3210	**0.3384**	2.02	2.0889	2.1800	2.0906
4*	0.3239	0.3313	0.3191	**0.3286**	2.02	2.1222	2.2000	2.1226
5*	0.3239	0.3385	0.3155	**0.3357**	2.02	2.1000	2.2100	2.0962
6*	0.3239	0.3275	0.3090	**0.3213**	2.02	2.1333	2.1400	2.1322
7	0.3239	observed	observed	observed	2.02	observed	observed	observed
8*	0.3239	0.3438	0.3290	**0.3457**	2.02	1.9778	1.8600	1.9772
9*	0.3239	0.3380	0.3290	**0.3317**	2.02	1.9778	1.8600	1.9788
10*	0.3239	0.3465	0.3259	**0.3395**	2.02	1.9667	1.8500	1.9644

MAR LNR High	**Standard Errors**			
Variable	CD	CC	SI	MI
1	0.0816	0.0816	0.0816	0.0816
2	0.0816	0.0837	0.0809	0.0830
3	0.0816	0.0860	0.0809	0.0863
4	0.0816	0.0834	0.0804	0.0838
5	0.0816	0.0852	0.0795	0.0856
6	0.0816	0.0824	0.0779	0.0819
7	0.0816	0.0816	0.0816	0.0816
8	0.0816	0.0865	0.0829	0.0882
9	0.0816	0.0851	0.0829	0.0846
10	0.0816	0.0872	0.0821	0.0866

CD – complete data, CC – complete-case analysis, SI – single imputation, MI – multiple imputation
* – indicates the variables with a wider confidence interval with regard to MI (only considering SI and MI)

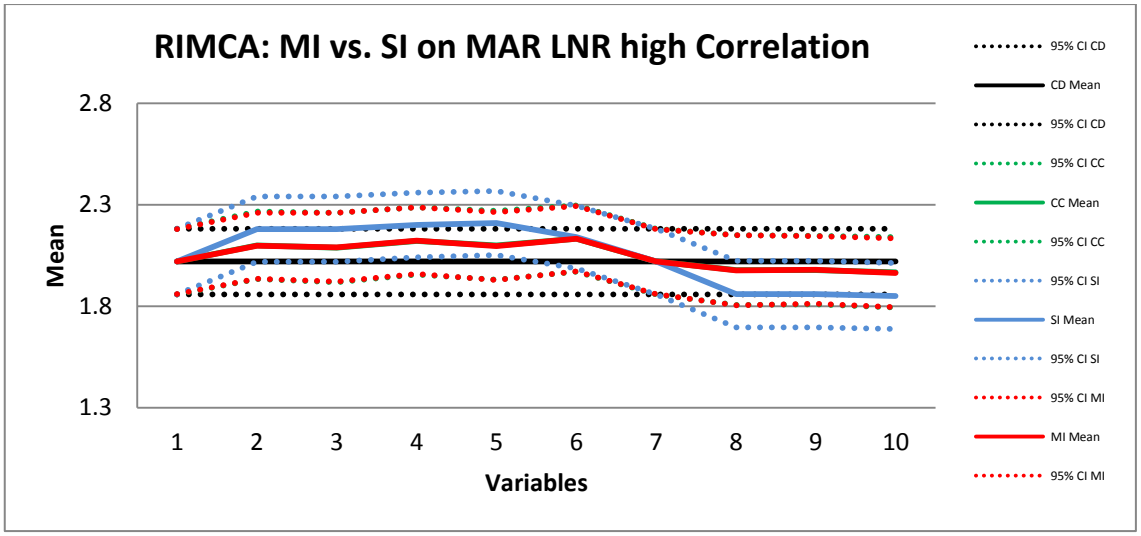


Figure 6.26 Means and Confidence intervals for RIMCA in MI and SI (MAR LNR)

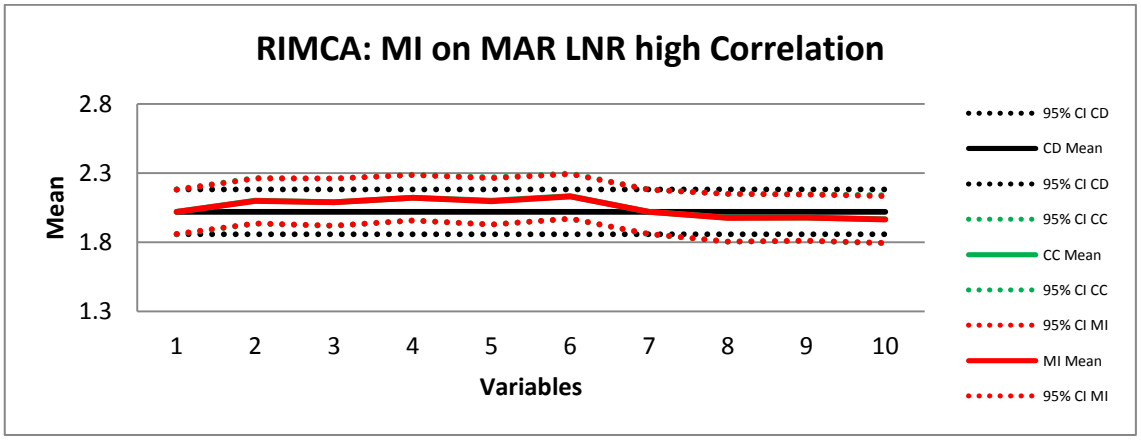


Figure 6.27 MI and CC vs. CD Mean and CI's on MAR LNR High correlated data

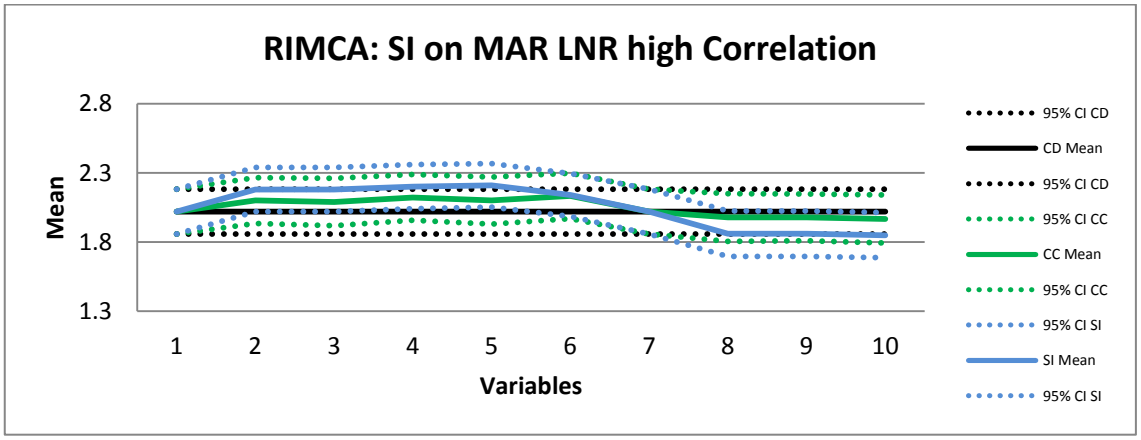


Figure 6.28 SI and CC vs. CD Mean and CI's on MAR LNR High correlated data

Discussion

Table 6.7 indicates that MI produces wider confidence intervals than SI for all variables. The MI confidence interval for variable three is slightly wider than the CC confidence interval. Further, all CC confidence intervals are wider than the MI confidence intervals. The graphical representation of this information is provided in figures 6.26 and 6.27. The mean estimates obtained from MI (**cf**. figure 6.27) for all of the imputed variables are closer to the true mean values than SI (**cf**. figure 6.28). Thus it is clear that MI is a better fit for the MAR high correlated data with a non-random pattern and low percentage of missing values than SI.

6.5.1.8 MAR LNR Low correlation structure

Table 6.8 Confidence interval widths, means and standard errors obtained from complete-case analysis and RIMCA in SI and MI for MAR LNR low correlated data in comparison to the true values

MAR LNR Low	Confidence Interval Width				Mean			
Variable	CD	CC	SI	MI	CD	CC	SI	MI
1	0.3239	observed	observed	observed	2.02	observed	observed	observed
2*	0.3239	0.3370	0.3275	**0.3354**	2.02	2.0667	2.1600	2.0580
3*	0.3239	0.3428	0.3051	**0.3364**	2.02	2.0667	2.0700	2.0634
4*	0.3239	0.3491	0.3195	**0.3429**	2.02	2.0444	2.0900	2.0418
5*	0.3239	0.3376	0.3109	**0.3325**	2.02	2.0444	2.0500	2.0436
6*	0.3239	0.3434	0.3290	**0.3361**	2.02	2.0444	2.1400	2.0458
7	0.3239	observed	observed	observed	2.02	observed	observed	observed
8*	0.3239	0.3352	0.3063	**0.3246**	2.02	1.9889	1.9900	1.9874
9*	0.3239	0.3380	0.3332	**0.3335**	2.02	1.9778	1.8900	1.9726
10	0.3239	0.3321	0.3332	0.3200	2.02	1.9778	1.8900	1.9794

MAR LNR Low	Standard Errors			
Variable	CD	CC	SI	MI
1	0.0816	0.0816	0.0816	0.0816
2	0.0816	0.0848	0.0825	0.0855
3	0.0816	0.0863	0.0769	0.0858
4	0.0816	0.0879	0.0805	0.0875
5	0.0816	0.0850	0.0783	0.0848
6	0.0816	0.0864	0.0829	0.0857
7	0.0816	0.0816	0.0816	0.0816
8	0.0816	0.0843	0.0772	0.0828
9	0.0816	0.0851	0.0840	0.0851
10	0.0816	0.0836	0.0840	0.0816

CD – complete data, CC – complete-case analysis, SI – single imputation, MI – multiple imputation
* – indicates the variables with a wider confidence interval with regard to MI (only considering SI and MI)

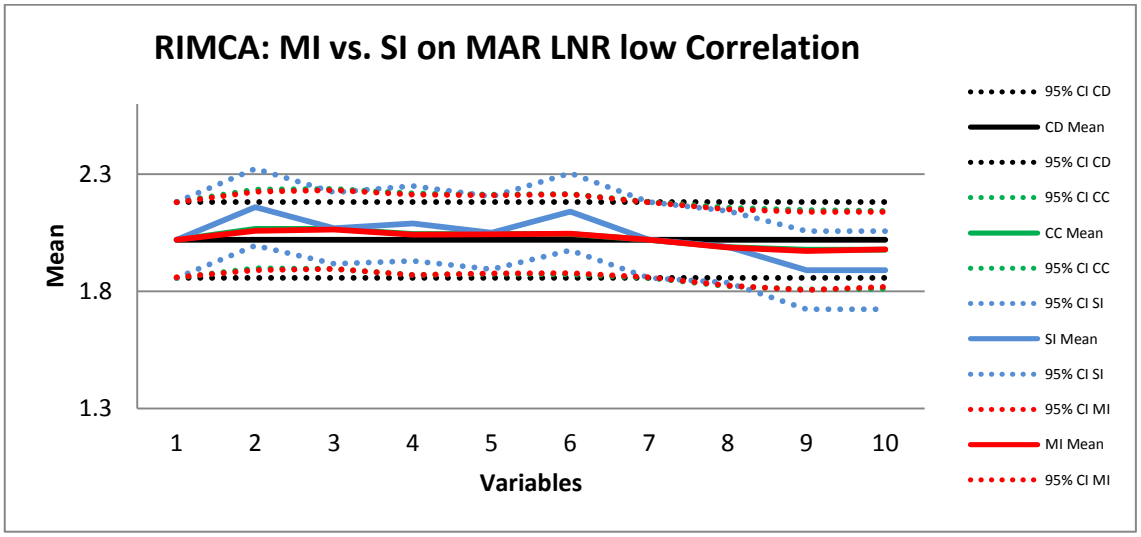


Figure 6.29 Means and Confidence intervals for RIMCA in MI and SI (MAR LNR)

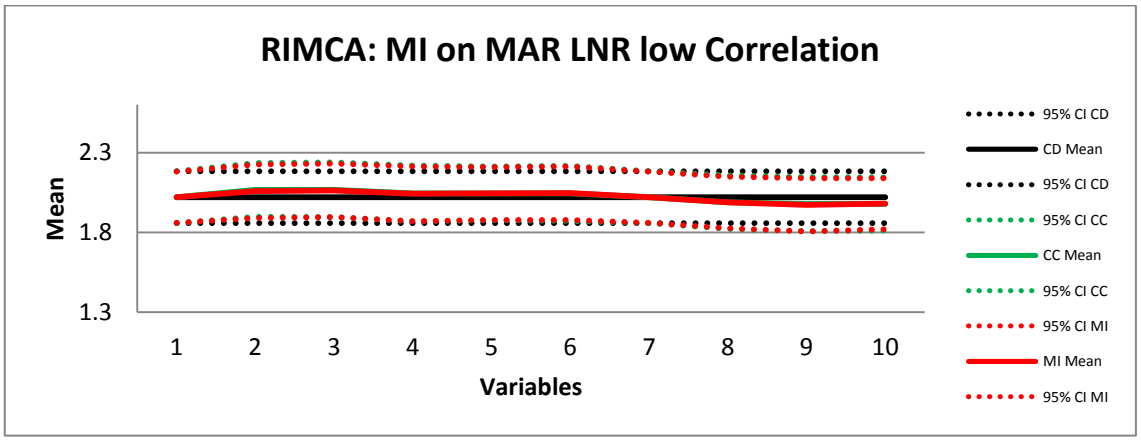


Figure 6.30 MI and CC vs. CD Mean and CI's on MAR LNR Low correlated data

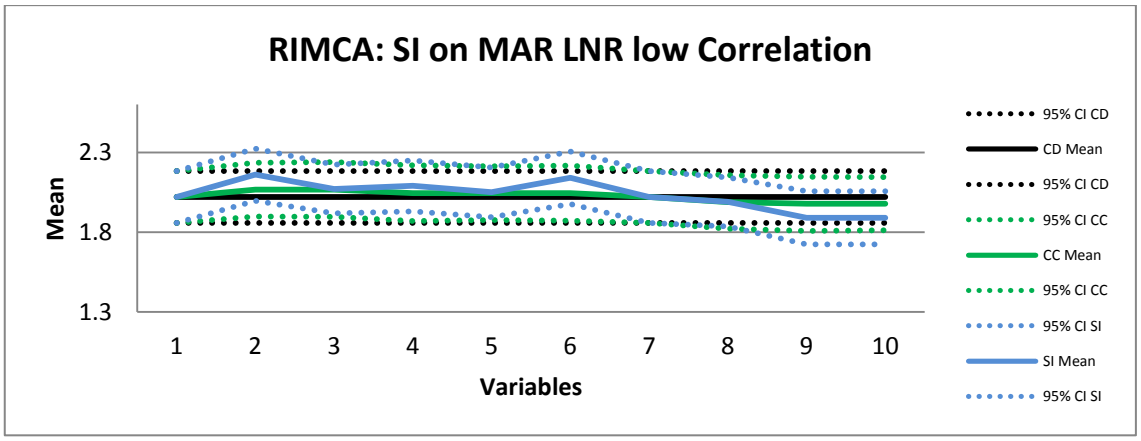


Figure 6.31 SI and CC vs. CD Mean and CI's on MAR LNR Low correlated data

Discussion

Table 6.8 indicates that MI produces wider confidence intervals than SI for all variables, except variable ten. The graphical representation of this information is provided in figure 6.29. The mean estimates obtained from MI for all of the imputed variables are closer to the true mean values (CD) than SI. From figures 6.29 and 6.30 it can be observed that MI fits the data well and provides better estimates than the SI (**cf**. figures 6.29 & 6.31) procedure. Once again the CC estimates closely resemble the MI estimates.

Conclusion for MAR LNR for High- and Low correlated data

As was the case with the preceding simulated datasets, MI performs slightly better in data with a low correlation structure. The estimates obtained from SI for the low correlation structure are closer to the true values than the estimates obtained from the high correlation structure from SI. In a majority of instances the confidence intervals obtained from MI are wider than the confidence intervals provided by SI.

Thus RIMCA in MI is a better fit for the specified data.

6.5.2 Simulated data with a MCAR missingness mechanism

The datasets with a MCAR missingness mechanism either have a low percentage (L) of missing values (10%) or a high percentage (H) of missing values (30%). The missing values are entered using a random (R) or non-random pattern (NR). Estimates are obtained for all of the variables in the case of the random pattern, but for the non-random pattern variables four, five, six, seven and eight are completely observed. Therefore, the confidence intervals of the SI and MI procedures will be equal to the true intervals for these observed variables. The analysis will be restricted to imputed variables only.

6.5.2.1 MCAR HR High correlation structure

Table 6.9 Confidence interval widths, means and standard errors obtained from complete-case analysis and RIMCA in SI and MI for MCAR HR high correlated data in comparison to the true values

MCAR HR High	Confidence Interval Width				Mean			
Variable	CD	CC	SI	MI	CD	CC	SI	MI
1*	0.3239	0.3967	0.3209	**0.3706**	2.02	1.9853	2.3500	1.9938
2*	0.3239	0.3783	0.3355	**0.3459**	2.02	2.0676	1.7500	2.0708
3*	0.3239	0.4045	0.2644	**0.3695**	2.02	2	1.9800	1.9944
4*	0.3239	0.3862	0.3489	**0.3553**	2.02	2.0278	2.0700	2.0244
5*	0.3239	0.3835	0.3274	**0.3608**	2.02	2.0282	2.3100	2.0338
6*	0.3239	0.4083	0.2912	**0.3881**	2.02	2.0968	1.8700	2.0986
7*	0.3239	0.3923	0.3225	**0.3674**	2.02	1.9589	2.3100	1.9686
8	0.3239	0.3864	0.3633	0.3553	2.02	2.0704	1.9900	2.0634
9*	0.3239	0.3628	0.3202	**0.3442**	2.02	2	2.3400	1.9978
10*	0.3239	0.3923	0.3084	**0.3551**	2.02	2	2.3900	1.9956

MCAR HR High	Standard Errors			
Variable	CD	CC	SI	MI
1	0.0816	0.0994	0.0809	0.0944
2	0.0816	0.0949	0.0845	0.0882
3	0.0816	0.1013	0.0666	0.0941
4	0.0816	0.0968	0.0879	0.0906
5	0.0816	0.0961	0.0825	0.0919
6	0.0816	0.1021	0.0734	0.0987
7	0.0816	0.0984	0.0813	0.0936
8	0.0816	0.0969	0.0916	0.0906
9	0.0816	0.0910	0.0807	0.0877
10	0.0816	0.0983	0.0777	0.0905

CD – complete data, CC – complete-case analysis, SI – single imputation, MI – multiple imputation
* – indicates the variables with a wider confidence interval with regard to MI (only considering SI and MI)

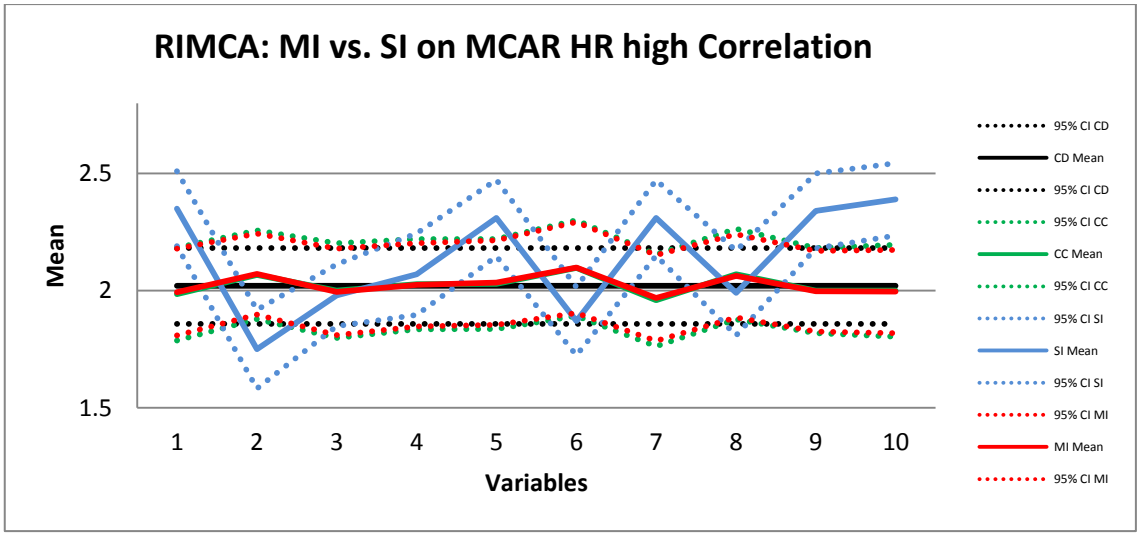


Figure 6.32 Means and Confidence intervals for RIMCA in MI and SI (MCAR HR)

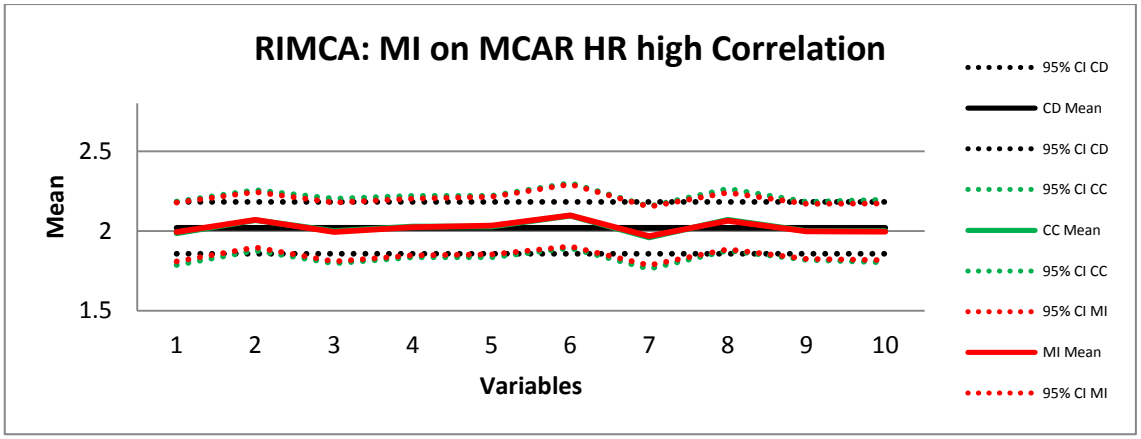


Figure 6.33 MI and CC vs. CD Mean and CI's on MCAR HR High correlated data

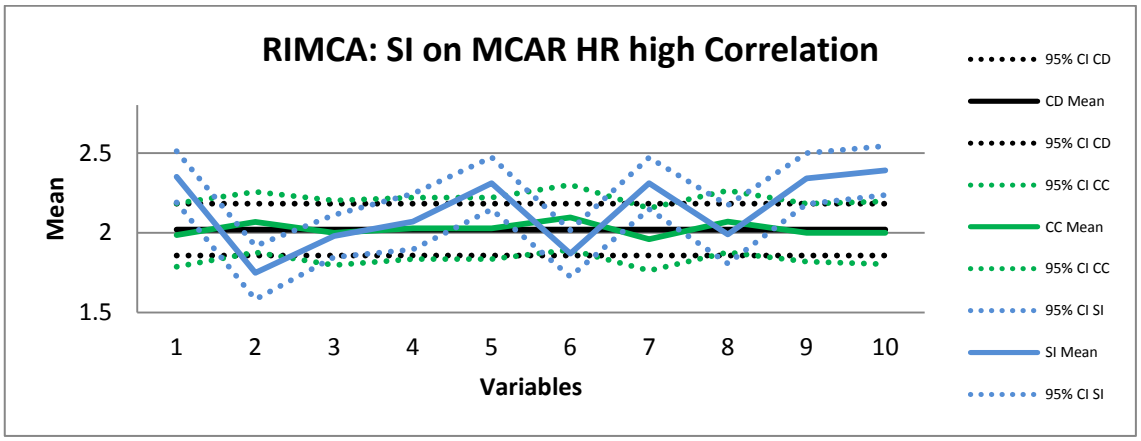


Figure 6.34 SI and CC vs. CD Mean and CI's on MCAR HR High correlated data

Discussion

Table 6.9 indicates that MI produces wider confidence intervals than SI for all variables. Further it is observed that the confidence intervals obtained from CC are wider than those of MI for all of the variables. The standard errors for the CC analysis are larger than those of SI and MI. The graphical representation of the means and confidence intervals is provided in figure 6.32. Figures 6.32 and 6.34 show that there is strong evidence that SI is statistically different from the true confidence intervals for variable ten, since the intervals for this variable do not overlap the true intervals. Therefore, for this specified variable SI will not provide accurate predictions. The mean estimates obtained from SI for variables three and eight are slightly closer to the true mean than the MI estimates. However, for all the remaining variables the mean estimates

provided by MI (**cf**. figure 6.33) are closer to the true mean than SI. The mean estimates obtained from CC are close to the mean estimates obtained from MI.

From the given information and discussion it is clear that MI is a better fit for the MCAR high correlated data with a random pattern and high percentage of missing values than SI.

6.5.2.2 MCAR HR Low correlation structure

Table 6.10 Confidence interval widths, means and standard errors obtained from complete-case analysis and RIMCA in SI and MI for MCAR HR low correlated data in comparison to the true values

MCAR HR Low	Confidence Interval Width				Mean			
Variable	CD	CC	SI	MI	CD	CC	SI	MI
1*	0.3239	0.3936	0.3180	**0.3609**	2.02	2.1029	1.6200	2.0918
2*	0.3239	0.4076	0.3173	**0.3742**	2.02	2.0313	1.6300	2.0278
3*	0.3239	0.3835	0.3251	**0.3467**	2.02	2.0282	1.6600	2.0280
4*	0.3239	0.3770	0.3021	**0.3663**	2.02	1.9863	2.1900	1.9772
5	0.3239	0.3606	0.3539	0.3392	2.02	1.9863	1.9500	1.9878
6*	0.3239	0.3740	0.3498	**0.3560**	2.02	2.0400	1.9700	2.0460
7*	0.3239	0.3821	0.2937	**0.3520**	2.02	1.9000	2.2400	1.9082
8*	0.3239	0.3911	0.3259	**0.3695**	2.02	1.9437	1.6500	1.9440
9*	0.3239	0.3836	0.3343	**0.3694**	2.02	1.9286	1.7600	1.9178
10*	0.3239	0.4201	0.3322	**0.3809**	2.02	2.0000	2.3100	2.0140

MCAR HR Low	Standard Errors			
Variable	CD	CC	SI	MI
1	0.0816	0.0986	0.0801	0.0920
2	0.0816	0.1020	0.0800	0.0953
3	0.0816	0.0961	0.0819	0.0884
4	0.0816	0.0945	0.0761	0.0933
5	0.0816	0.0904	0.0892	0.0864
6	0.0816	0.0938	0.0881	0.0907
7	0.0816	0.0958	0.0740	0.0897
8	0.0816	0.0980	0.0821	0.0941
9	0.0816	0.0961	0.0842	0.0941
10	0.0816	0.1052	0.0837	0.0970

CD – complete data, CC – complete-case analysis, SI – single imputation, MI – multiple imputation
* – indicates the variables with a wider confidence interval with regard to MI (only considering SI and MI)

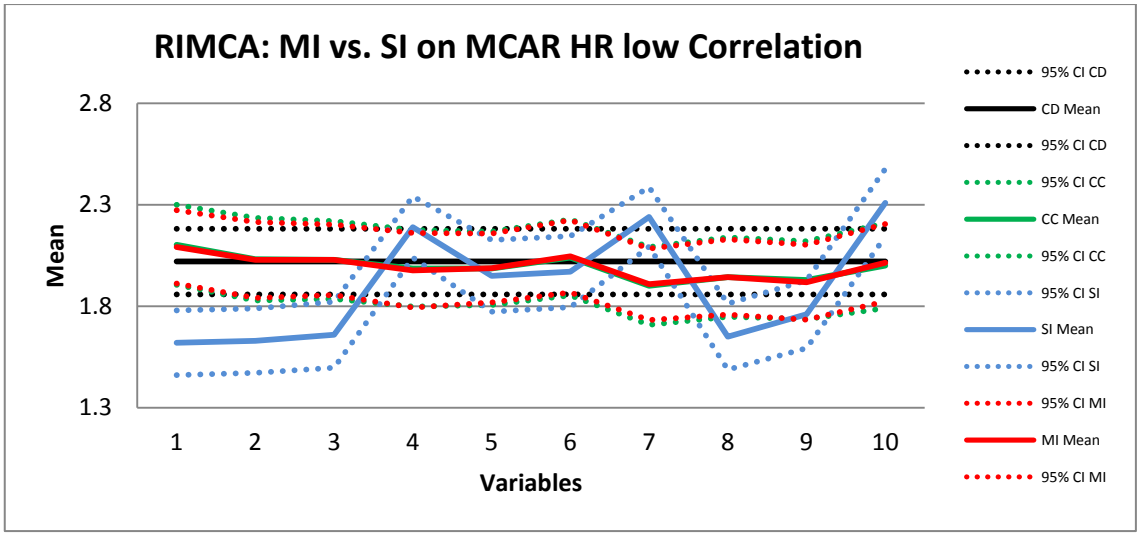


Figure 6.35 Means and Confidence intervals for RIMCA in MI and SI (MCAR HR)

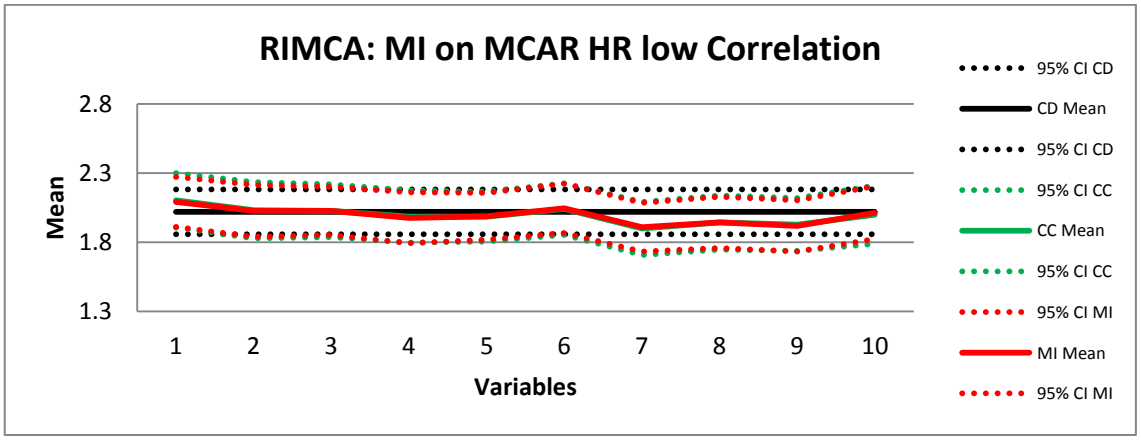


Figure 6.36 MI and CC vs. CD Mean and CI's on MCAR HR Low correlated data

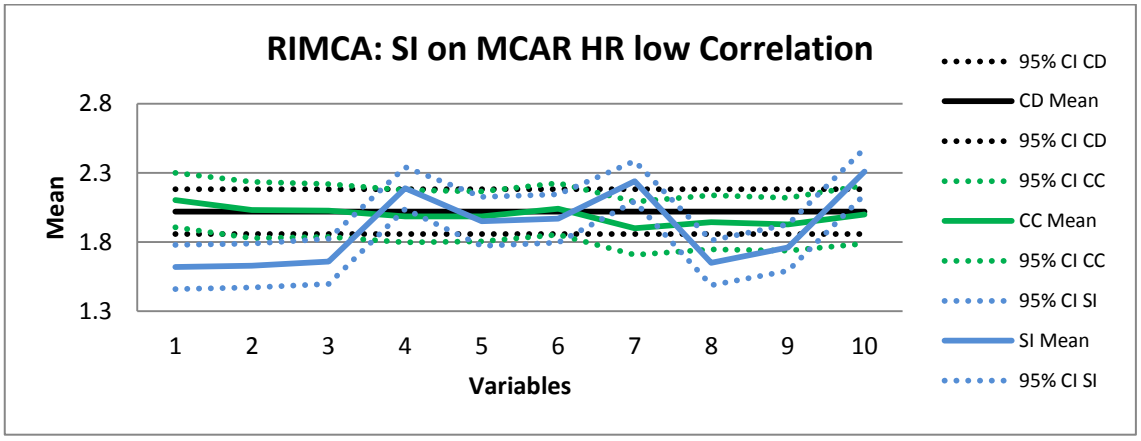


Figure 6.37 SI and CC vs. CD Mean and CI's on MCAR HR Low correlated data

Discussion

Table 6.10 indicates that MI produces wider confidence intervals than SI for all variables except variable five, with CC providing even wider confidence intervals than MI. The graphical representation of this information is provided in figure 6.35. Figures 6.35 and 6.37 show that there is strong evidence that SI is statistically different from the true confidence intervals for variables one, two, three and eight, since the intervals for these variables do not overlap the true intervals. Therefore, for this specified variable SI will not provide accurate predictions. The mean estimates of CC and MI are again highly correlated and are closer to the true mean values than the SI estimates for all variables (**cf**. figure 6.35).

Thus it is clear that MI is a better fit for the MCAR low correlated data with a random pattern and high percentage of missing values than SI.

Conclusion for MCAR HR for High- and Low correlated data

MI performs better in the high correlated data with respect to the mean estimates. SI also performs better in the high correlated data, since four of variables in the low correlation case are statistically different from the true mean and confidence intervals. In both correlation cases MI outperforms SI; providing more accurate mean estimates and wider confidence intervals.

6.5.2.3 MCAR HNR High correlation structure

Table 6.11 Confidence interval widths, means and standard errors obtained from complete-case analysis and RIMCA in SI and MI for MCAR HNR high correlated data in comparison to the true values

MCAR HNR High	Confidence Interval Width				Mean			
Variable	CD	CC	SI	MI	CD	CC	SI	MI
1*	0.3239	0.5458	0.2801	**0.4345**	2.02	2.2000	2.6300	2.2206
2*	0.3239	0.5119	0.3770	**0.3921**	2.02	2.0250	2.0800	2.0232
3*	0.3239	0.5572	0.1871	**0.4220**	2.02	2.1000	2	2.1044
4	0.3239	observed	observed	observed	2.02	observed	observed	observed
5	0.3239	observed	observed	observed	2.02	observed	observed	observed
6	0.3239	observed	observed	observed	2.02	observed	observed	observed
7	0.3239	observed	observed	observed	2.02	observed	observed	observed
8	0.3239	observed	observed	observed	2.02	observed	observed	observed
9*	0.3239	0.5098	0.2677	**0.4046**	2.02	1.9250	2.6400	1.9212
10*	0.3239	0.5300	0.1864	**0.4094**	2.02	1.9250	2.0400	1.9160

MCAR HNR High	Standard Errors			
Variable	CD	CC	SI	MI
1	0.0816	0.1349	0.0706	0.1104
2	0.0816	0.1265	0.0950	0.0997
3	0.0816	0.1377	0.0471	0.1073
4	0.0816	0.0816	0.0816	0.0816
5	0.0816	0.0816	0.0816	0.0816
6	0.0816	0.0816	0.0816	0.0816
7	0.0816	0.0816	0.0816	0.0816
8	0.0816	0.0816	0.0816	0.0816
9	0.0816	0.1260	0.0674	0.1028
10	0.0816	0.1310	0.0470	0.1041

CD – complete data, CC – complete-case analysis, SI – single imputation, MI – multiple imputation
* – indicates the variables with a wider confidence interval with regard to MI (only considering SI and MI)

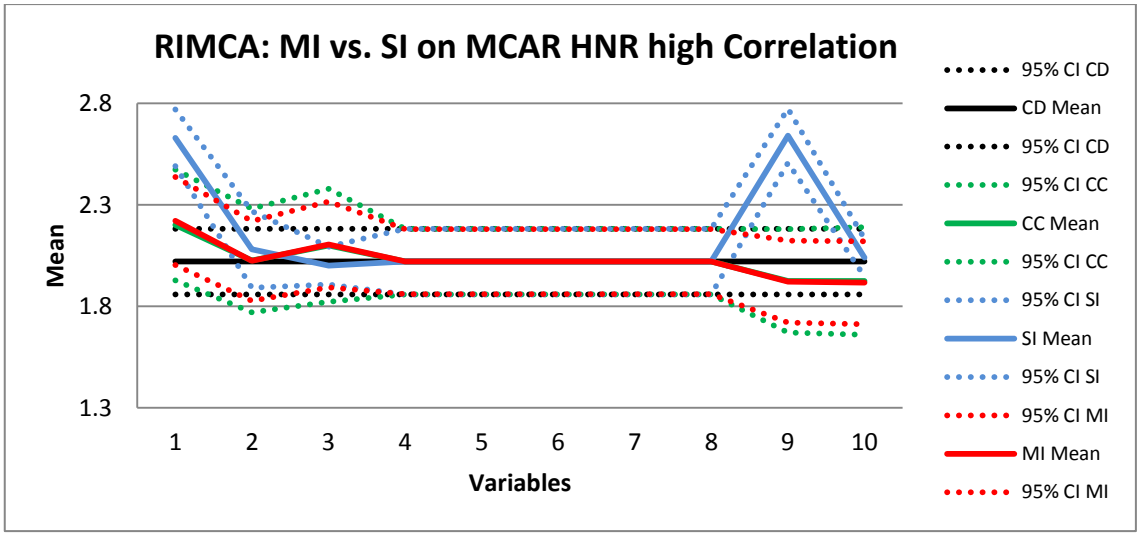


Figure 6.38 Means and Confidence intervals for RIMCA in MI and SI (MCAR HNR)

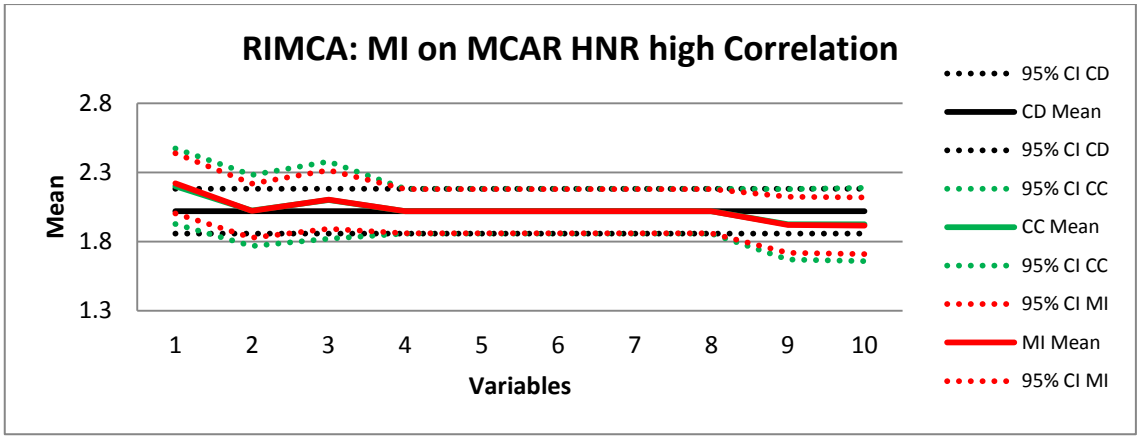


Figure 6.39 MI and CC vs. CD Mean and CI's on MCAR HNR High correlated data

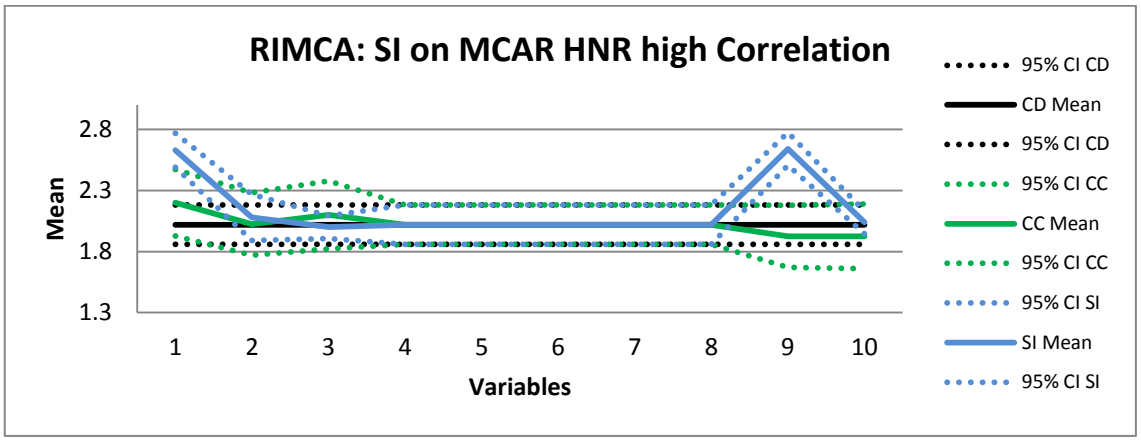


Figure 6.40 SI and CC vs. CD Mean and CI's on MCAR HNR High correlated data

Discussion

Table 6.11 indicates that MI produces wider confidence intervals than SI for all variables. The graphical representation of this information is provided in figure 6.38. Figures 6.38 and 6.40 show that there is strong evidence that SI is statistically different from the true confidence intervals for variables one and nine, since the intervals for these variables do not overlap the true intervals. Therefore, SI is not applicable in these instances and does not provide sufficient results. The mean estimate obtained from SI for variable three and ten is closer to the true mean than the MI estimate. However, for all the remaining variables the mean estimates provided by MI (**cf**. figure 6.39) are closer to the true mean than SI. The CC estimates are similar to the MI estimates and provide wider confidence intervals than the MI procedure.

6.5.2.4 MCAR HNR Low correlation structure

Table 6.12 Confidence interval widths, means and standard errors obtained from complete-case analysis and RIMCA in SI and MI for MCAR HNR low correlated data in comparison to the true values

MCAR HNR Low	Confidence Interval Width				Mean			
Variable	CD	CC	SI	MI	CD	CC	SI	MI
1*	0.3239	0.5300	0.1952	**0.3962**	2.02	1.9250	2.0200	1.9304
2*	0.3239	0.5456	0.2735	**0.4290**	2.02	2.1250	1.3600	2.1220
3*	0.3239	0.5213	0.3738	**0.3892**	2.02	2.0500	1.9600	2.0302
4	0.3239	observed	observed	observed	2.02	observed	observed	observed
5	0.3239	observed	observed	observed	2.02	observed	observed	observed
6	0.3239	observed	observed	observed	2.02	observed	observed	observed
7	0.3239	observed	observed	observed	2.02	observed	observed	observed
8	0.3239	observed	observed	observed	2.02	observed	observed	observed
9*	0.3239	0.5119	0.1994	**0.3858**	2.02	2.0250	2.0100	2.0216
10*	0.3239	0.5420	0.2814	**0.4344**	2.02	2	2.6100	2.0046

MCAR HNR Low	Standard Errors			
Variable	CD	CC	SI	MI
1	0.0816	0.1310	0.0492	0.1008
2	0.0816	0.1349	0.0689	0.1090
3	0.0816	0.1289	0.0942	0.0990
4	0.0816	0.0816	0.0816	0.0816
5	0.0816	0.0816	0.0816	0.0816
6	0.0816	0.0816	0.0816	0.0816
7	0.0816	0.0816	0.0816	0.0816
8	0.0816	0.0816	0.0816	0.0816
9	0.0816	0.1265	0.0502	0.0981
10	0.0816	0.1340	0.0709	0.1104

CD – complete data, CC – complete-case analysis, SI – single imputation, MI – multiple imputation
* – indicates the variables with a wider confidence interval with regard to MI (only considering SI and MI)

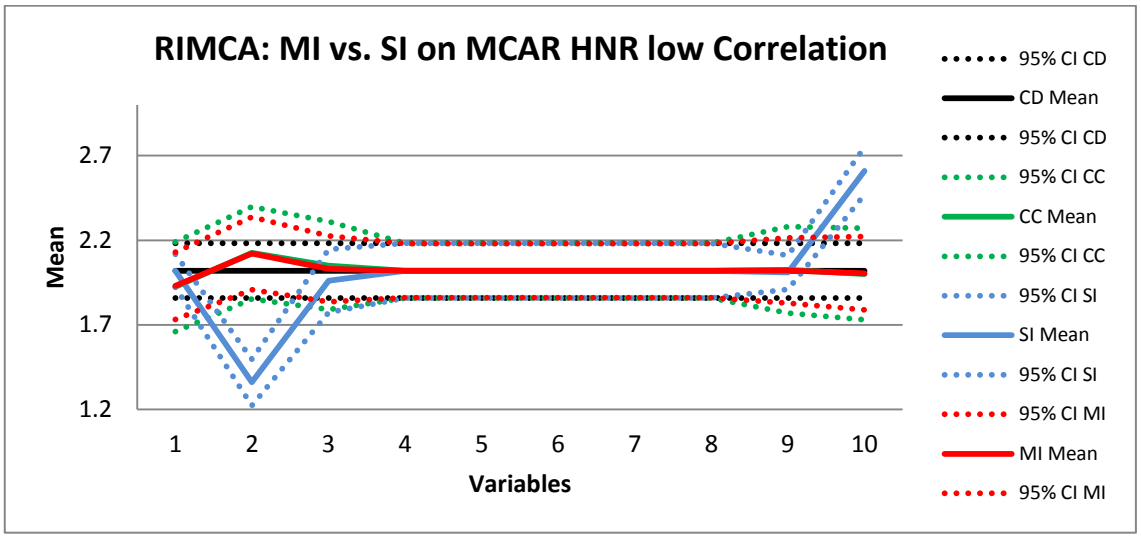


Figure 6.41 Means and Confidence intervals for RIMCA in MI and SI (MCAR HNR)

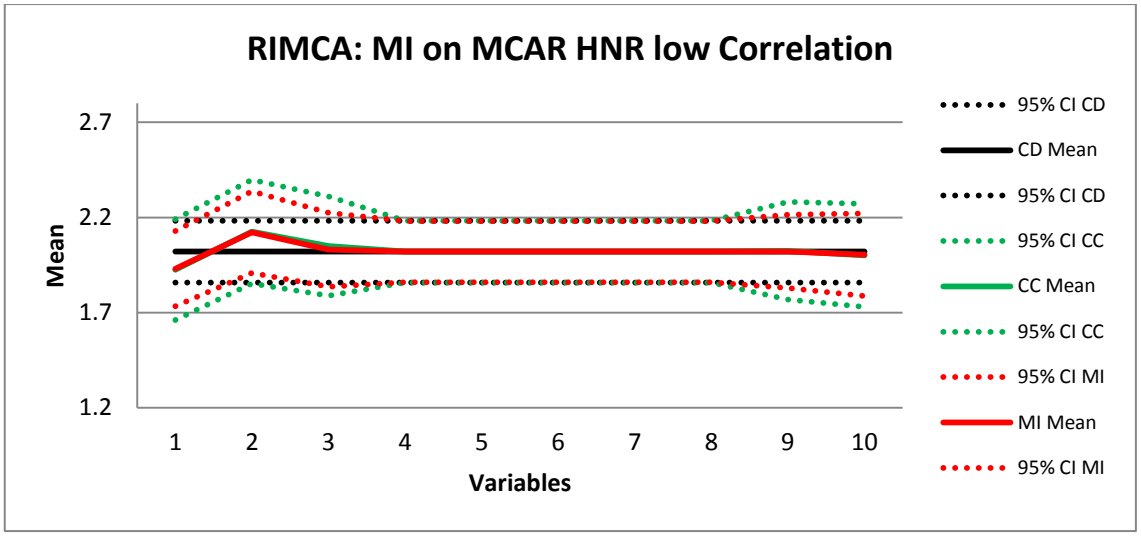


Figure 6.42 MI and CC vs. CD Mean and CI's on MCAR HNR Low correlated data

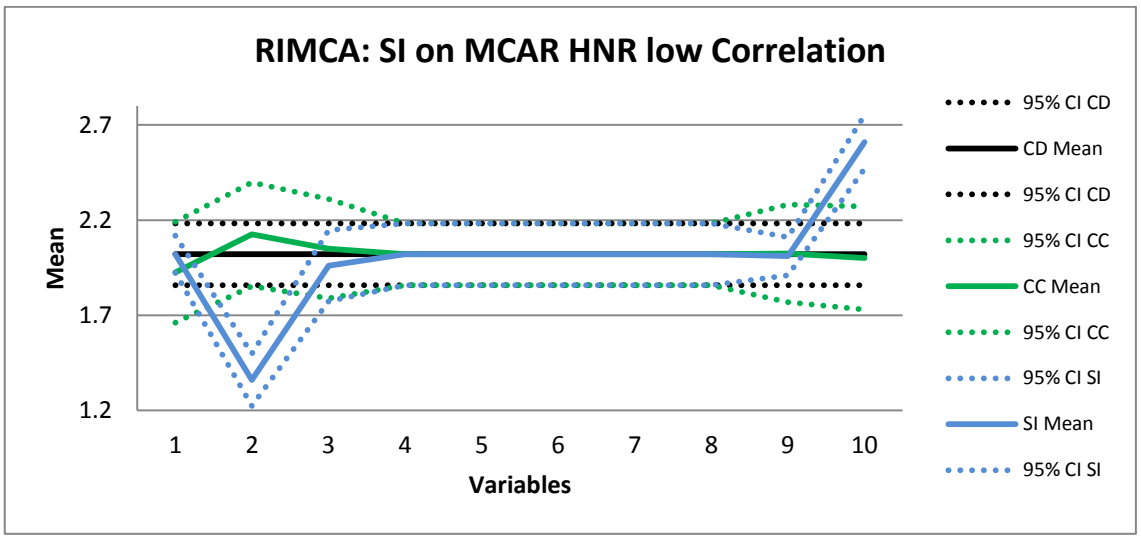


Figure 6.43 SI and CC vs. CD Mean and CI's on MCAR HNR Low correlated data

Discussion

Table 6.12 indicates that MI produces wider confidence intervals than SI for all variables. It can also be observed that CC provides wider confidence intervals than MI, with regard to the mean estimates these two approaches provide similar results. The graphical representation of this information is provided in figure 6.41. Figures 6.41 and 6.43 show that there is strong evidence that SI is statistically different from the true confidence intervals for variables two and ten, since the intervals for these variables do not overlap the true intervals. Therefore, SI is not applicable in these instances and does not provide sufficient results. The mean estimate obtained from SI for variable one is equal

to the true mean (2.02). However, for all the remaining variables the mean estimates provided by MI (**cf**. figure 6.42) are closer to the true mean than SI.

Conclusion for MCAR HNR for High- and Low correlated data

Since there is strong evidence that the results obtained from SI are statistically different from the true values for two of the variables, it is not a good fit for this particular data. To the contrary MI performs well in both correlation structures, but slightly better in the presence of a low correlation structure. Again the added uncertainty is confirmed by the wider confidence intervals provided by MI.

6.5.2.5 MCAR LR High correlation structure

Table 6.13 Confidence interval widths, means and standard errors obtained from complete-case analysis and RIMCA in SI and MI for MCAR LR high correlated data in comparison to the true values

MCAR LR High	Confidence Interval Width				Mean			
Variable	CD	CC	SI	MI	CD	CC	SI	MI
1*	0.3239	0.3434	0.2958	**0.3444**	2.02	2.0460	1.9900	2.0390
2*	0.3239	0.3421	0.3223	**0.3345**	2.02	2.0714	2.1300	2.0690
3*	0.3239	0.3460	0.3381	**0.3430**	2.02	2	2.0400	2.0014
4*	0.3239	0.3474	0.3259	**0.3363**	2.02	2.0449	2.1500	2.0446
5*	0.3239	0.3351	0.3063	**0.3305**	2.02	2.0323	2.0100	2.0364
6*	0.3239	0.3391	0.2958	**0.3320**	2.02	1.9891	1.9900	1.9914
7*	0.3239	0.3358	0.3061	**0.3304**	2.02	2.0435	2.0300	2.0374
8	0.3239	0.3396	0.3302	0.3299	2.02	2.0440	2.1200	2.0480
9	0.3239	0.3363	0.3423	0.3288	2.02	2	2.0600	2.0016
10*	0.3239	0.3307	0.3231	**0.3249**	2.02	2	2.0600	1.9988

MCAR LR High	Standard Errors			
Variable	CD	CC	SI	MI
1	0.0816	0.0864	0.0745	0.0878
2	0.0816	0.0860	0.0812	0.0853
3	0.0816	0.0870	0.0852	0.0875
4	0.0816	0.0874	0.0821	0.0858
5	0.0816	0.0844	0.0772	0.0843
6	0.0816	0.0854	0.0745	0.0847
7	0.0816	0.0845	0.0771	0.0843
8	0.0816	0.0855	0.0832	0.0841
9	0.0816	0.0847	0.0862	0.0839
10	0.0816	0.0832	0.0814	0.0829

CD – complete data, CC – complete-case analysis, SI – single imputation, MI – multiple imputation
* – indicates the variables with a wider confidence interval with regard to MI (only considering SI and MI)

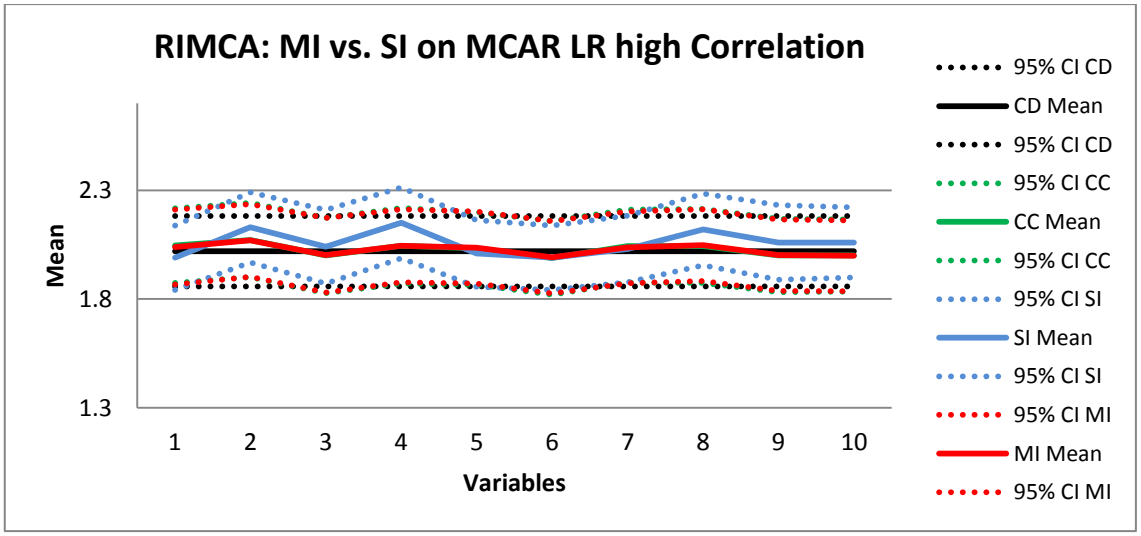


Figure 6.44 Means and Confidence intervals for RIMCA in MI and SI (MCAR LR)

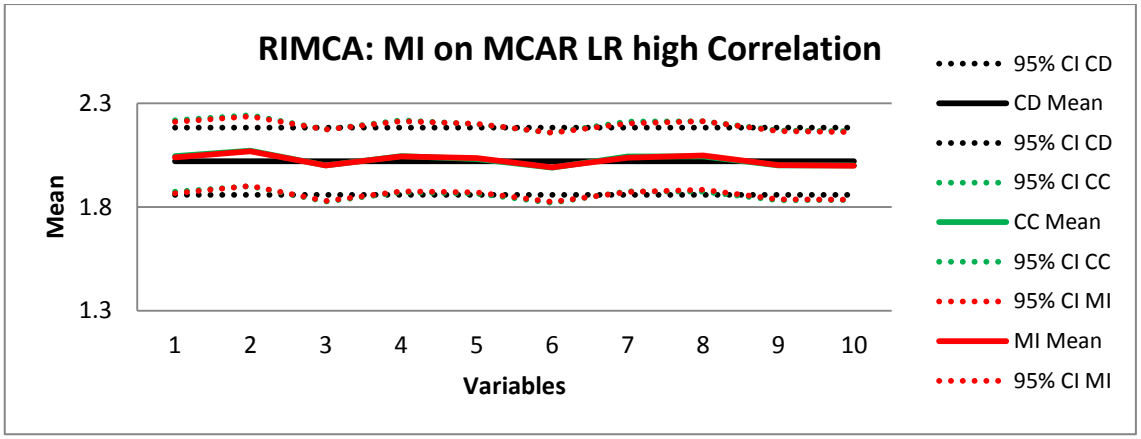


Figure 6.45 MI and CC vs. CD Mean and CI's on MCAR LR High correlated data

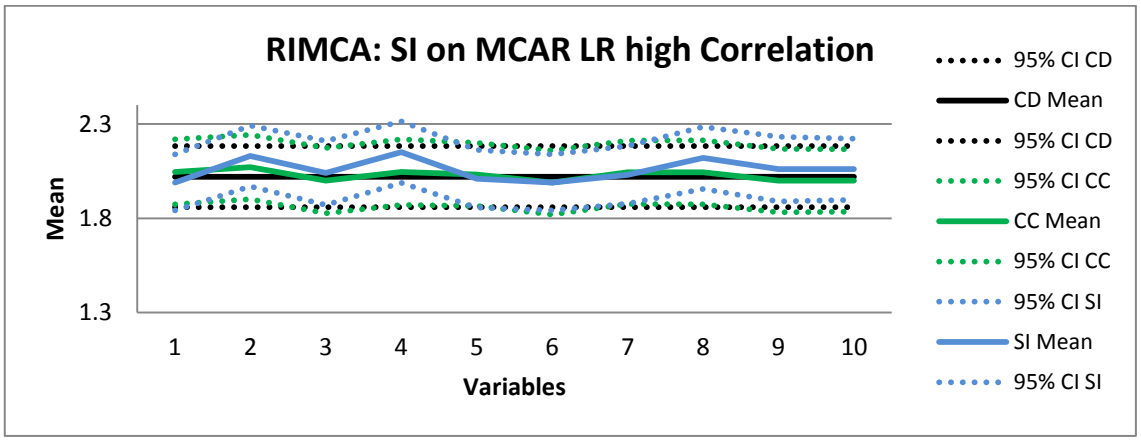


Figure 6.46 SI and CC vs. CD Mean and CI's on MCAR LR High correlated data

Discussion

Table 6.13 indicates that MI produces wider confidence intervals than SI for all variables, except variables eight and nine. Further the confidence intervals for CC are larger than the confidence intervals obtained from MI for all variables, except variable one. The graphical representation of this information is provided in figure 6.44. The mean estimates obtained from SI for variables five and seven are slightly closer to the true mean value in comparison to the mean estimates of MI. Regardless of this, the mean estimates for all the remaining variables provided by MI (**cf**. figure 6.45) are closer to the true mean than SI (**cf**. figure 6.46). The standard errors obtained from the CC analysis are greater than the standard errors obtained from the imputation procedures in a majority of the variables. Again the MI estimates and CC estimates are similar.

6.5.2.6 MCAR LR Low correlation structure

Table 6.14 Confidence interval widths, means and standard errors obtained from complete-case analysis and RIMCA in SI and MI for MCAR LR low correlated data in comparison to the true values

MCAR LR Low	Confidence Interval Width				Mean			
Variable	CD	CC	SI	MI	CD	CC	SI	MI
1*	0.3239	0.3391	0.3290	**0.3346**	2.02	1.9891	2.1400	1.9956
2*	0.3239	0.3460	0.3313	**0.3320**	2.02	2.0556	2.100	2.0550
3*	0.3239	0.3441	0.2984	**0.3405**	2.02	2.0562	1.9800	2.0628
4*	0.3239	0.3381	0.3322	**0.3339**	2.02	2	2.0800	2.0002
5*	0.3239	0.3487	0.3338	**0.3424**	2.02	2.0341	2.1400	2.0318
6*	0.3239	0.3408	0.2998	**0.3335**	2.02	2.0674	2.0700	2.0648
7*	0.3239	0.3455	0.3236	**0.3448**	2.02	2.0455	1.9600	2.0402
8*	0.3239	0.3400	0.3195	**0.3326**	2.02	1.9780	2.0900	1.9790
9*	0.3239	0.3530	0.2958	**0.3520**	2.02	1.9885	2.0100	1.9860
10	0.3239	0.3351	0.3313	0.3305	2.02	1.9677	1.9000	1.9720

MCAR LR Low	Standard Errors			
Variable	CD	CC	SI	MI
1	0.0816	0.0854	0.0829	0.0854
2	0.0816	0.0871	0.0835	0.0847
3	0.0816	0.0866	0.0752	0.0868
4	0.0816	0.0851	0.0837	0.0852
5	0.0816	0.0877	0.0841	0.0873
6	0.0816	0.0858	0.0756	0.0850
7	0.0816	0.0869	0.0816	0.0879
8	0.0816	0.0856	0.0805	0.0848
9	0.0816	0.0888	0.0745	0.0897
10	0.0816	0.0844	0.0835	0.0843

CD – complete data, CC – complete-case analysis, SI – single imputation, MI – multiple imputation
* – indicates the variables with a wider confidence interval with regard to MI (only considering SI and MI)

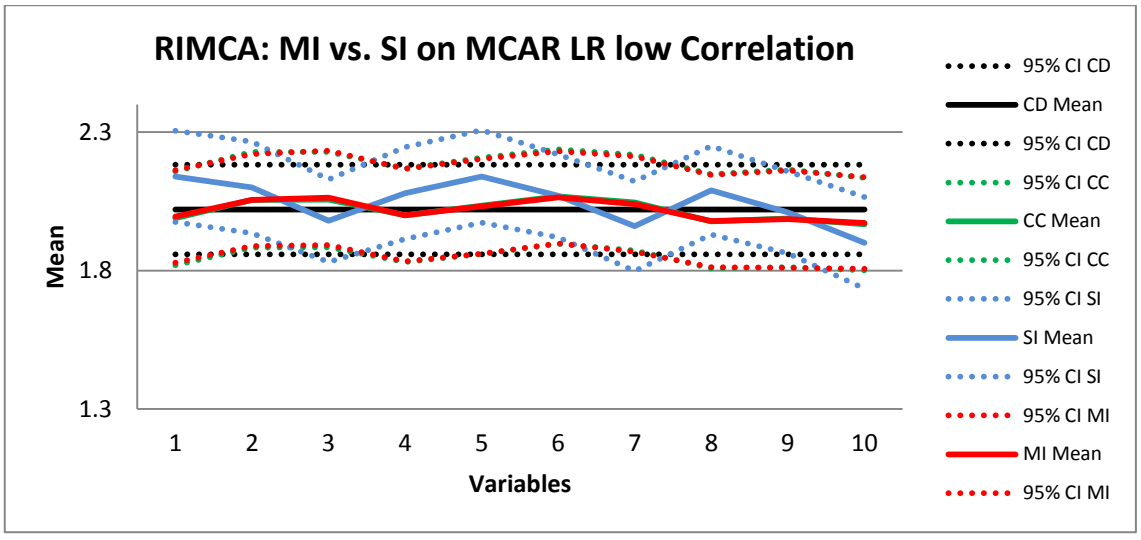


Figure 6.47 Means and Confidence intervals for RIMCA in MI and SI (MCAR LR)

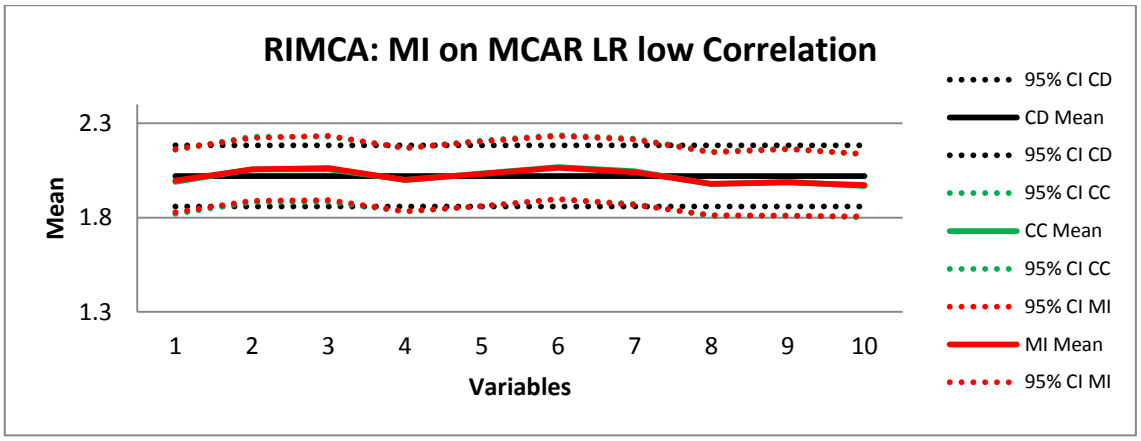


Figure 6.48 MI and CC vs. CD Mean and CI's on MCAR LR Low correlated data

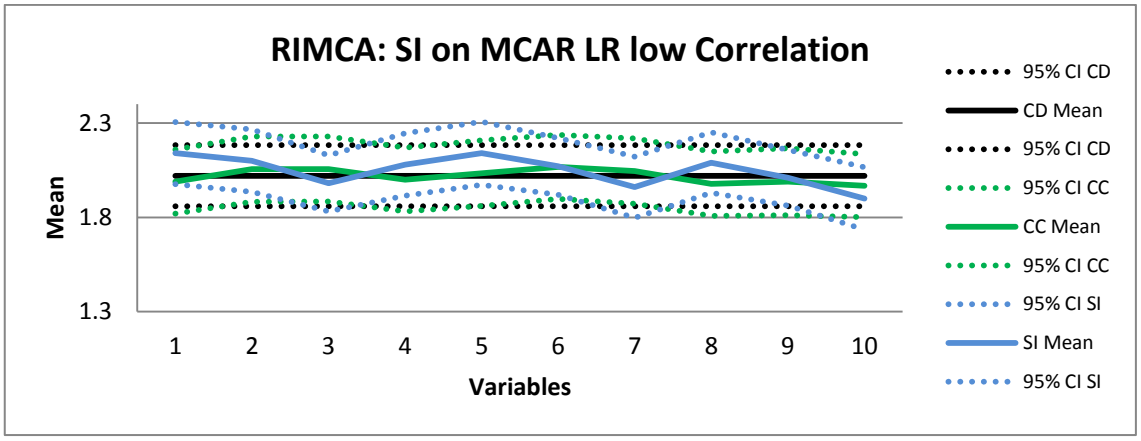


Figure 6.49 SI and CC vs. CD Mean and CI's on MCAR LR Low correlated data

Discussion

Table 6.14 indicates that MI produces wider confidence intervals for all variables, except variable ten. The graphical representation of this information is provided in figure 6.47. The mean estimates obtained from SI (**cf**. figure 6.49) for variables three and nine are slightly closer to the true mean value. However, for all the remaining variables the mean estimates provided by MI are closer to the true mean than SI. MI and CC provide highly correlated estimates, with CC providing slightly wider confidence intervals than MI. Also, the standard errors obtained from CC are larger than the standard errors obtained from SI and MI. From figures 6.47 and 6.48 it can be observed that MI fits the true values well with mean estimates closely resembling the true mean values (CD mean).

Conclusion for MCAR LR for High- and Low correlated data

The performance of both imputation procedures is slightly better in the high correlated data. RIMCA as a MI procedure fits the data well and incorporates wider confidence intervals in a majority of the variables. This confirms that RIMCA in MI outperforms RIMCA in SI for this particular data.

6.5.2.7 MCAR LNR High correlation structure

Table 6.15 Confidence interval widths, means and standard errors obtained from complete-case analysis and from RIMCA in SI and MI for MCAR LNR high correlated data in comparison to the true values

MCAR LNR High	Confidence Interval Width				Mean			
Variable	CD	CC	SI	MI	CD	CC	SI	MI
1*	0.3239	0.3710	0.3519	**0.3544**	2.02	2.0375	2.0400	2.0428
2*	0.3239	0.3572	0.2875	**0.3435**	2.02	2.0375	1.9800	2.0380
3*	0.3239	0.3747	0.3351	**0.3617**	2.02	2	1.7900	1.9980
4	0.3239	observed	observed	observed	2.02	observed	observed	observed
5	0.3239	observed	observed	observed	2.02	observed	observed	observed
6	0.3239	observed	observed	observed	2.02	observed	observed	observed
7	0.3239	observed	observed	observed	2.02	observed	observed	observed
8	0.3239	observed	observed	observed	2.02	observed	observed	observed
9*	0.3239	0.3611	0.3282	**0.3385**	2.02	2	2.2300	1.9952
10*	0.3239	0.3576	0.2875	**0.3489**	2.02	2.0125	2.0200	2.0114

MCAR LNR High	Standard Errors			
Variable	CD	CC	SI	MI
1	0.0816	0.0932	0.0887	0.0904
2	0.0816	0.0897	0.0724	0.0876
3	0.0816	0.0941	0.0844	0.0922
4	0.0816	0.0816	0.0816	0.0816
5	0.0816	0.0816	0.0816	0.0816
6	0.0816	0.0816	0.0816	0.0816
7	0.0816	0.0816	0.0816	0.0816
8	0.0816	0.0816	0.0816	0.0816
9	0.0816	0.0907	0.0827	0.0863
10	0.0816	0.0898	0.0724	0.0889

CD – complete data, CC – complete-case analysis, SI – single imputation, MI – multiple imputation
* – indicates the variables with a wider confidence interval with regard to MI (only considering SI and MI)

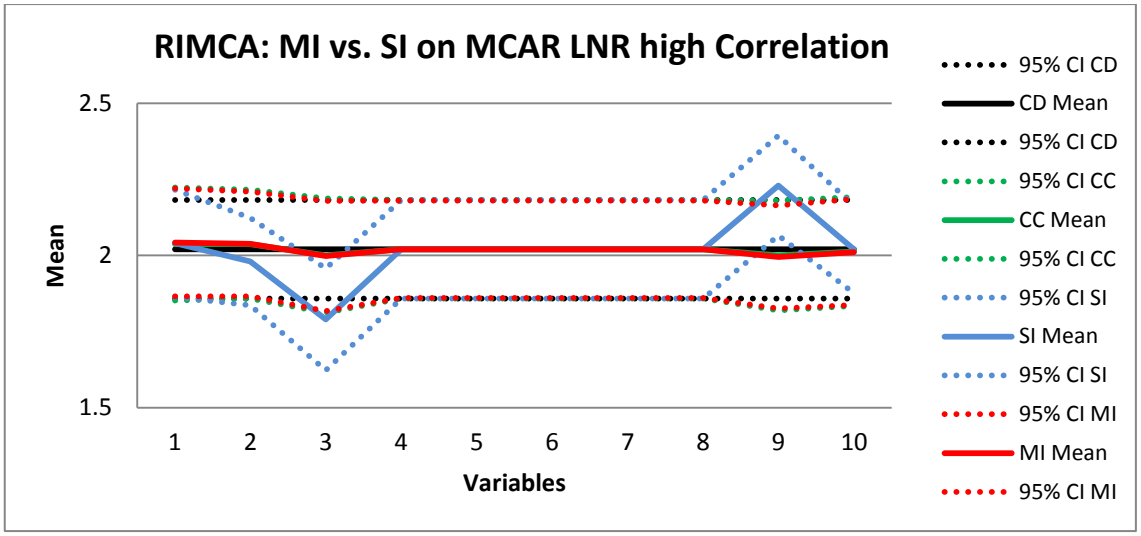


Figure 6.50 Means and Confidence intervals for RIMCA in MI and SI (MCAR LNR)

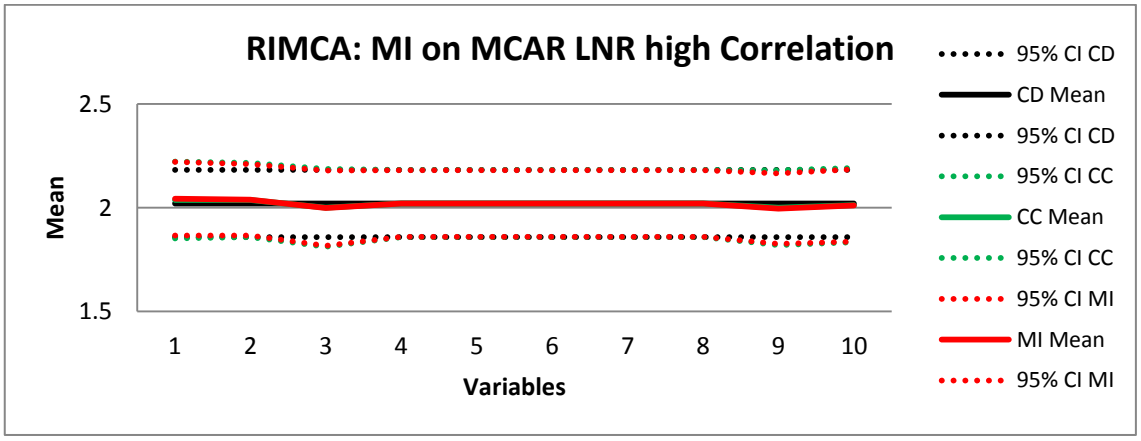


Figure 6.51 MI and CC vs. CD Mean and CI's on MCAR LNR High correlated data

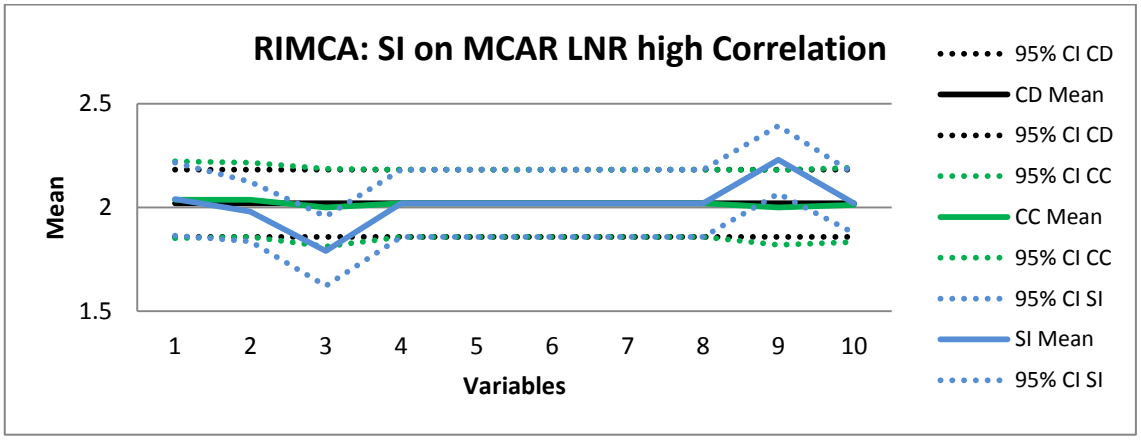


Figure 6.52 SI and CC vs. CD Mean and CI's on MCAR LNR High correlated data

Discussion

Table 6.15 indicates that MI produces wider confidence intervals for all variables, with CC providing even wider confidence intervals for all variables. The graphical representation of this information is provided in figure 6.50. The mean estimate obtained from SI (**cf**. figure 6.52) for variable one is slightly closer to the true mean value (CD mean) than the MI estimate and for variable ten the SI estimate is equal to the true mean. Regardless of this, for all the remaining variables the mean estimates provided by MI are closer to the true mean than SI. The CC and MI estimates are similar (**cf**. figure 6.51). From figures 6.50 and 6.52 it can be observed that the MI mean estimates closely resemble the true mean values and shows good fit.

6.5.2.8 MCAR LNR Low correlation structure

Table 6.16 Confidence interval widths, means and standard errors obtained from complete-case analysis and RIMCA in SI and MI for MCAR LNR low correlated data in comparison to the true values

MCAR LNR Low	Confidence Interval Width				Mean			
Variable	CD	CC	SI	MI	CD	CC	SI	MI
1*	0.3239	0.3673	0.2876	**0.3484**	2.02	2.0500	2	2.0556
2*	0.3239	0.3678	0.3282	**0.3546**	2.02	2.0250	1.7700	2.0280
3*	0.3239	0.3678	0.2848	**0.3580**	2.02	1.9750	2.0100	1.9698
4	0.3239	observed	observed	observed	2.02	observed	observed	observed
5	0.3239	observed	observed	observed	2.02	observed	observed	observed
6	0.3239	observed	observed	observed	2.02	observed	observed	observed
7	0.3239	observed	observed	observed	2.02	observed	observed	observed
8	0.3239	observed	observed	observed	2.02	observed	observed	observed
9*	0.3239	0.3611	0.2848	**0.3443**	2.02	2	2.0100	2.0026
10*	0.3239	0.3680	0.3173	**0.3551**	2.02	2	2.1300	1.9834

MCAR LNR Low	Standard Errors			
Variable	CD	CC	SI	MI
1	0.0816	0.0923	0.0725	0.0888
2	0.0816	0.0924	0.0827	0.0904
3	0.0816	0.0924	0.0718	0.0913
4	0.0816	0.0816	0.0816	0.0816
5	0.0816	0.0816	0.0816	0.0816
6	0.0816	0.0816	0.0816	0.0816
7	0.0816	0.0816	0.0816	0.0816
8	0.0816	0.0816	0.0816	0.0816
9	0.0816	0.0907	0.0718	0.0878
10	0.0816	0.0924	0.0800	0.0905

CD – complete data, CC – complete-case analysis, SI – single imputation, MI – multiple imputation
* – indicates the variables with a wider confidence interval with regard to MI (only considering SI and MI)

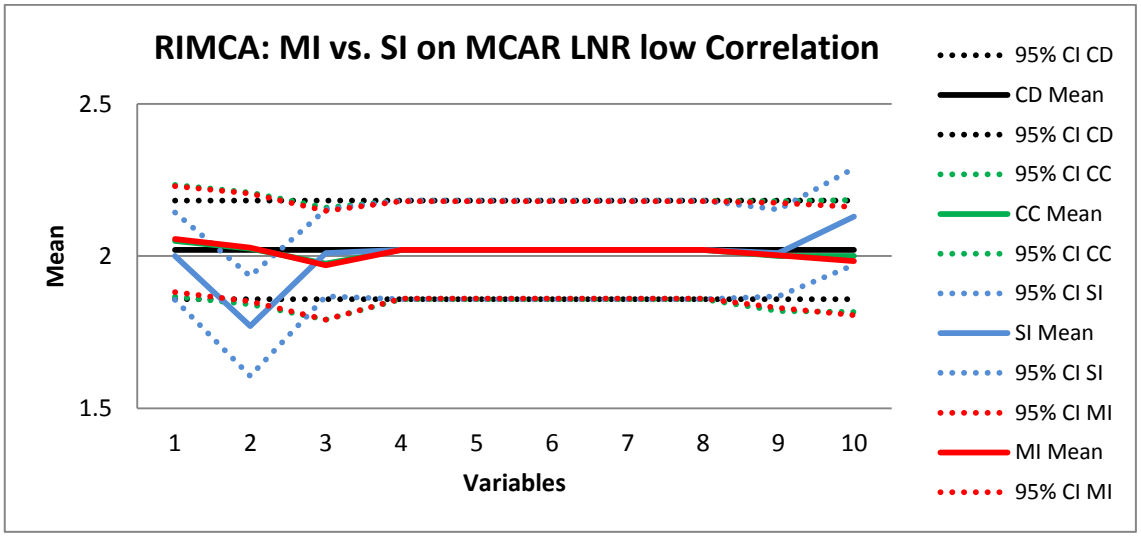


Figure 6.53 Means and Confidence intervals for RIMCA in MI and SI (MCAR LNR)

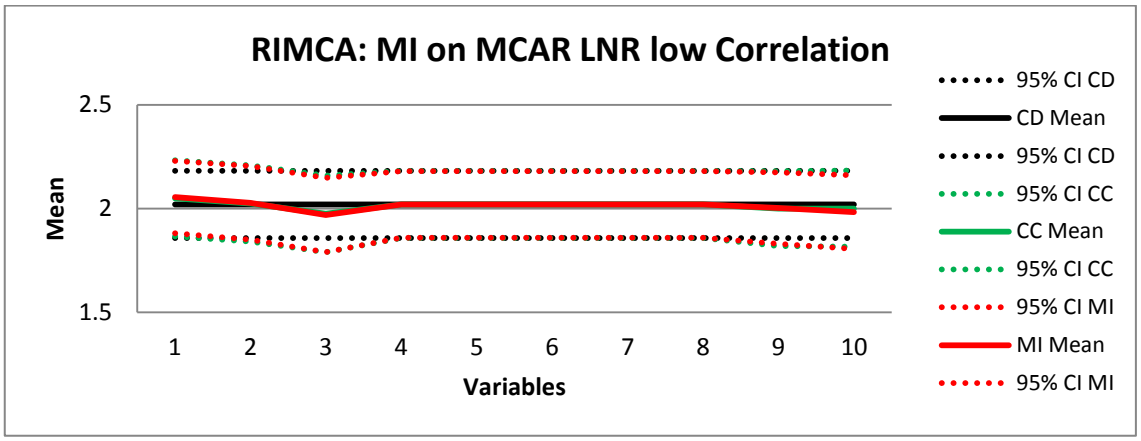


Figure 6.54 MI and CC vs. CD Mean and CI's on MCAR Low High correlated data

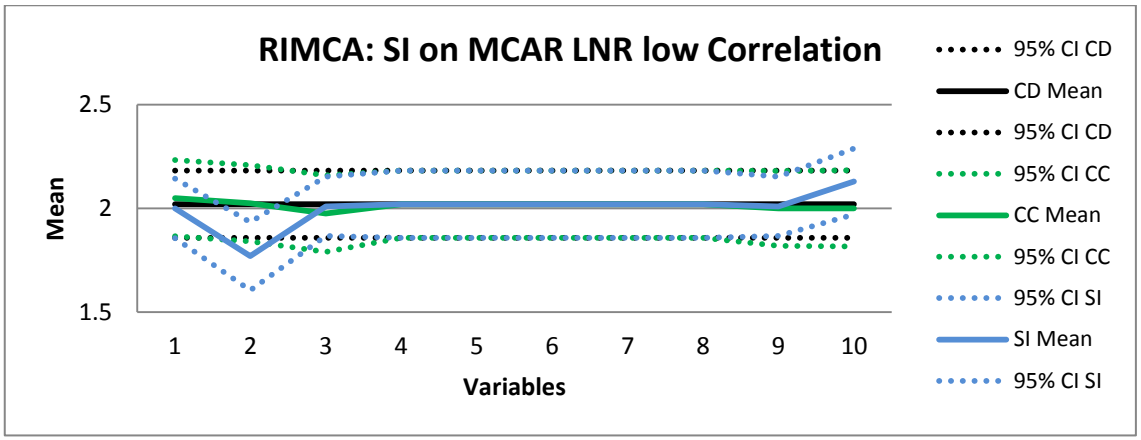


Figure 6.55 SI and CC vs. CD Mean and CI's on MCAR LNR Low correlated data

Discussion

Table 6.16 indicates that MI produces wider confidence intervals for all variables and the confidence intervals provided by CC are even wider than the MI intervals. The graphical representation of this information is provided in figure 6.53. The standard errors obtained from CC analysis are larger than the standard errors of the imputation procedures. The mean estimates obtained from RIMCA in MI (**cf**. figure 6.54) are closer to the true mean values for most of the variables than RIMCA in SI (**cf**. figure 6.55). The MI procedure shows good fit in this particular dataset.

Conclusion for MCAR LNR for High- and Low correlated data

SI and MI perform slightly better in the low correlation structure of the simulated data of this context. Once again the confidence intervals provided by MI are wider for all variables, as well as estimates showing better fit. Therefore MI outperforms SI in this dataset.

6.5.3 Objective one: Conclusion

Rubin's (1987:21) statement regarding the degrees of freedom obtained for the calculation of the confidence intervals is confirmed. Large values for the degrees of freedom (v) were to be expected in the presence of small between variance (B_m) values with respect to the total variance (T). Another indication would be the use of a large number of multiple datasets (m). Both of these indications are observed by the results of the generated data, in each case the between-variance (B_m) values are smaller than the total variance (T) and ten multiple datasets are generated, which is considered as a large number (Rubin 1987:2, 15 & **cf**. 3.4.3.2). In the presence of large degrees of freedom values, the t-critical values will be Normal (1.96) with regard to a 95% confidence interval. The tables of the statistics obtained from Rubin's rules (**cf**. 3.5) are available in appendix L.

MAR mechanism

The confidence intervals obtained from MI are wider in each of the variables for each of the simulated datasets with a MAR mechanism. This confirms that the uncertainty added by MI is successfully incorporated. MI performs slightly better in data with a low correlation structure. This is interesting, since literature (Josse *et al.* 2012:114) states that RIMCA in SI is expected to perform better for highly correlated data and MAR values. It is found that the mean estimates provided by MI are closer to the true means than the SI estimates. Also, in the MAR HR and MAR HNR high correlated data SI provided estimates that are statistically different from the true values. RIMCA in MI is a better fit for the data and incorporates the additional uncertainty expected from MI.

MCAR mechanism

In all of the datasets MI provides better mean estimates and wider confidence intervals in most of the variables. In the MCAR HR and MCAR HNR datasets for both correlation structures SI provided inaccurate estimates for some of the variables which are statistically different from the true values. The difference in the performance of RIMCA in MI with regard to the correlation structures is very slight. MI performs well in both correlation structures.

In general

For all simulated datasets it is found that the CC estimates and MI estimates are extremely close to each other, therefore the assumption can be made that estimates close to the estimates obtained from CC analysis will provide satisfying results in the context of the specific data. The MI confidence intervals are narrower than the confidence intervals provided by CC analysis. The narrow confidence intervals can be explained by perhaps a small amount of variance that is added on each imputation. An underestimated variance was expected, because of the fixed parameter values (**cf**. 5.5).

Even though one of the great advantages of MI is the increase in the uncertainty (**cf**. 3.4.3.2), Rubin (1996:482) argues that from a statistical

viewpoint sharper confidence intervals with at least 95% coverage will be preferred over wider confidence intervals with exactly 95% coverage. Rubin (1996:481) also discuss the occurrence of superefficient imputations, which could be explained by the allowance of extra information in the sense of allowing for extra uncertainty, as opposed to single imputation methods and complete-case analysis, therefore the estimates provided by MI could be more efficient and precise than the estimates provided by CC.

The advantages (Rubin 1978:16) of MI over SI listed in the literature review of MI (**cf**. 3.4.3.2) were confirmed by the results.

It can be seen that in each case the MI estimates closely resemble the true means in comparison to the SI estimates. Albeit slight, the estimates obtained from RIMCA in SI are more variable in some cases than those obtained from MI. This confirms the advantage stating that MI increases the accuracy of estimation. Another advantage is that in combining the multiple datasets valid inferences are attained incorporating additional variance caused by the missing values.

In the literature extreme variability can be a result of an imputation model that does not fit the data well and is far from the actual data or where an extreme uncertainty occurs regarding the missingness mechanism (Meng 1994:555). Since all missingness mechanisms were known, this confirms that in most cases SI does not fit the actual data well.

This confirms that RIMCA in MI performs well and is a suitable imputation technique.

6.6 Objective two: To investigate the accuracy of the predictions made by RIMCA in MI when applied to a simulated dataset.

The results obtained for this objective will be presented by means of tables consisting of the apparent error rates (incorrect imputations) and overall success rates (correct imputations). Information regarding the performance of RIMCA in MI and RIMCA in SI will be provided and compared. In order to determine the accuracy of the predictions obtained from the RIMCA algorithm in MI, apparent error rates are calculated for each of the S randomly selected dimensions (**cf**. 5.3.2) as was used in objective one. Since only one dimension is chosen *a priori* for SI, the apparent error rates for all possible S dimensions are used to establish the apparent error rates for SI. This will guide the reader to allocate the dimensions which performs better for single imputation purposes.

It is important to note that the high percentage of missing values in the MAR mechanisms are 16% and for the low percentage of missing values is 8%. The high percentage of missing values present in the MCAR datasets is 30% and the low percentage of missing values is 10%. Therefore the average values presented in the tables show the percentage of incorrect imputations made with respect to the percentage of missing values for the specific dataset. The success rates for the imputations with regard to the percentage of missing values are also given, finally displaying the overall percentage of correctly imputed values per specific dataset.

Table 6.17, 6.19, 6.21 and 6.23 displays the error rates for the data simulated over all dimensions, for the SI tables (**cf**. table 6.21 & 6.23) the highlighted cells indicate the smallest error rate which is achieved for the corresponding number of dimensions per dataset. Table 6.18, 6.20, 6.22 and 6.24 represent the average error rates obtained over all of the dimensions with respect to the percentage of missing values for the specific simulated dataset in question. The apparent success rates with respect to the percentage of missing values, along

with the overall success rates of the imputations are presented. The apparent error rates obtained from RIMCA in MI will be presented and discussed and followed by the results obtained from RIMCA in SI.

6.6.1 Apparent error rates: RIMCA in MI

The results obtained from a selection of ten dimensions are displayed in table 6.17 and 6.19 with the low and high correlation structures, respectively. The datasets obtained from each dimension are indicated by *MI 1, MI2, MI3,* etc. Also the error rates and success rates are presented in table 6.18 and 6.20.

Table 6.17 Apparent error rates: RIMCA in MI with a low correlation structure

Dataset	High percentage of missing values				Low percentage of missing values			
	Random		**Non-random**		**Random**		**Non-random**	
	MAR	MCAR	MAR	MCAR	MAR	MCAR	MAR	MCAR
	16%	30%	16%	30%	8%	10%	8%	10%
MI 1	10.16	20.20	10.70	20.40	5.70	6.70	5.24	6.86
MI 2	11.14	20.04	10.96	20.08	5.32	7	5.48	6.68
MI 3	10.90	20.86	10.82	20.06	5.32	6.58	5.54	6.62
MI 4	10.60	20.70	11.18	19.78	5.62	6.58	5.38	6.42
MI 5	11.04	20.12	11.30	20.54	5.62	6.60	5.38	6.82
MI 6	10.58	20.28	10.92	19.58	5.54	6.56	5.26	6.84
MI 7	11.28	20.14	11.28	19.86	5.78	6.56	5.54	6.80
MI 8	11.32	20.34	10.98	19.68	5.90	6.54	5.38	6.48
MI 9	10.58	19.82	11.40	20.32	5.62	6.66	5.40	6.70
MI 10	11.14	20.04	11.28	20.36	5.64	6.96	5.54	6.60

Table 6.18 Apparent error rates and success rates of the imputations made by RIMCA in MI for simulated data with a low correlation structure

RIMCA in MI Low correlated data	Average apparent error rate	Average apparent success rate	Average overall apparent success rate
MAR HR	**10.874 %** of 16%	**5.126 %** of 16%	32.0375 %
MCAR HR	**20.254 %** of 30%	**9.746 %** of 30%	32.49 %
MAR HNR	**11.082 %** of 16%	**4.918 %** of 16%	30.74 %
MCAR HNR	**20.066%** of 30%	**9.934 %** of 30%	33.11 %
MAR LR	**5.606 %** of 8%	**2.394 %** of 8%	29.925 %
MCAR LR	**6.674 %** of 10%	**3.326 %** of 10%	33.26 %
MAR LNR	**5.414 %** of 8%	**2.586 %** of 8%	32.325 %
MCAR LNR	**6.682 %** of 10%	**3.318 %** of 10%	33.18 %

The interpretation of table 6.18 will be done in depth, providing guidance for the interpretation of table 6.20, 6.22 and 6.24.

In the MAR data with a random pattern and high percentage of missing values, 10.874% of the possible 16% missing values are imputed incorrectly, which results in a success rate of 5.126% and an overall imputation success rate of 32.0375%. The MAR data with a non-random pattern and high percentage of missing values shows incorrect imputations are made 11.082% of the time, implying a success rate of 4.918% and an overall 30.74% of missing values being imputed correctly. The MCAR values with random and non-random patterns obtained success rates of 9.746% and 9.934%, respectively. The overall imputation success rate for MCAR values with a random pattern is 32.49% and for the non-random pattern is 33.11%. In the case of the low percentage of missing values present in the datasets, the MAR mechanism with a random pattern has a success rate of 2.394% with regard to the percentage of missing values, which results in an overall success rate of 29.925%. The MAR mechanism with a non-random pattern obtained an overall success rate of 32.325%. The MCAR mechanisms differ slightly with respect to patterns; the random pattern has an overall success rate of 32.26%, whereas the non-random pattern has a 33.18% overall success rate.

On average, irrespective of the simulation protocol followed per dataset, 32.13% of all the imputations made by RIMCA as a MI procedure are correct for the simulated data with a low correlation structure. The imputations made for the simulated data with MCAR values perform slightly better than the MAR cases, when taking the same pattern and percentage of missing values into consideration.

Table 6.19 Apparent error rates: RIMCA in MI with a high correlation structure

Dataset	High percentage of missing values				Low percentage of missing values			
	Random		**Non-random**		**Random**		**Non-random**	
	MAR	MCAR	MAR	MCAR	MAR	MCAR	MAR	MCAR
	16%	30%	16%	30%	8%	10%	8%	10%
MI 1	11.74	20.28	12.56	19.90	5.64	6.82	5.48	6.54
MI 2	12	20.90	12.14	19.34	5.60	6.38	5.66	6.90
MI 3	11.84	20	11.86	20.48	5.78	7.02	5.68	6.76
MI 4	11.92	20.04	11.62	20.20	5.62	6.92	5.70	6.74
MI 5	12.30	19.72	11.66	20.26	5.66	6.72	5.78	6.72
MI 6	12.32	20.22	11.62	20.18	5.78	6.62	5.56	6.84
MI 7	11.64	19.84	11.88	20.70	5.56	6.90	5.68	6.74
MI 8	12.32	19.22	11.90	19.98	5.94	6.34	5.68	6.66
MI 9	12.18	19.74	11.66	20.40	5.98	6.84	5.62	6.76
MI 10	12.20	20.38	12.50	20.22	5.74	7.06	5.86	6.64

Table 6.20 Apparent error rates and success rates of the imputations made by RIMCA in MI for simulated data with a high correlation structure

RIMCA in MI High correlated data	Average apparent error rate	Average apparent success rate	Average overall apparent success rate
MAR HR	**12.046 %** of 16%	**3.954 %** of 16%	24.71 %
MCAR HR	**20.034 %** of 30%	**9.966 %** of 30%	33.22 %
MAR HNR	**11.94 %** of 16%	**4.06 %** of 16%	25.375 %
MCAR HNR	**20.166 %** of 30%	**9.834 %** of 30%	32.78 %
MAR LR	**5.73 %** of 8%	**2.27 %** of 8%	28.375 %
MCAR LR	**6.762 %** of 10%	**3.238 %** of 10%	32.38 %
MAR LNR	**5.67 %** of 8%	**2.33 %** of 8%	29.125 %
MCAR LNR	**6.73 %** of 10%	**3.27 %** of 10%	32.7 %

On average, irrespective of the simulation protocol followed per dataset, 29.83% of all the imputations made by RIMCA as a MI procedure are correct for the simulated data with a high correlation structure. Again it can be observed that the imputations made for the simulated data with MCAR values perform better than the MAR cases when taking the same pattern and percentage of missing values into consideration.

Discussion

The apparent error rates for each of the multiple datasets for a specific missingness mechanism, percentage of missing values, specific pattern and correlation structure are similar. This once again confirms that the different dimensions perform similarly and that the RIMCA algorithm produces correlated

results for each of the dimensions once convergence is reached. The overall success rates were higher for the MAR cases in the presence of a low correlation structure. Whereas the success rates for the MCAR cases were higher than MAR for both high and low correlation structures.

It should be noted that the error rates are only an indication of the performance of the RIMCA algorithm in terms of the accuracy of the guesswork and not an indication of which dimensions would deliver the best results. The error rates only determine the accuracy of the RIMCA algorithm to predict missing values in the context of the simulated datasets.

6.6.2 Apparent error rates: RIMCA in SI

Again, the results obtained from all the possible S dimensions are displayed in table 6.21 and 6.23 with the low and high correlation structures, respectively. As well as the error rates and success rates illustrated in table 6.22 and 6.24.

Table 6.21 Apparent error rates: RIMCA in SI with a low correlation structure

# Dimensions	High percentage of missing values				Low percentage of missing values			
	Random		**Non-random**		**Random**		**Non-random**	
	MAR	MCAR	MAR	MCAR	MAR	MCAR	MAR	MCAR
	16%	30%	16%	30%	8%	10%	8%	10%
1	12.7	21.7	13.2	19.4	7.4	8.1	6.4	7.4
2	12.7	21.7	13.0	19.5	7.4	7.5	6.4	7.4
3	13.2	21.5	13.0	19.5	7.4	7.8	6.5	7.3
4	12.8	21.2	12.9	19.4	7.4	7.9	6.4	7.3
5	12.9	21.4	13.0	19.4	7.4	8.0	6.4	7.2
6	12.9	21.4	13.3	19.4	7.4	7.8	6.4	7.1
7	12.9	21.2	13.2	19.7	7.4	7.7	6.5	7.2
8	12.9	21.3	13.1	20.0	7.4	7.7	6.4	7.2
9	12.7	21.3	13.1	20.2	7.4	7.7	6.4	7.1
10	12.7	21.3	13.1	20.2	7.4	7.6	6.4	7.5
11	12.6	21.6	13.0	20.0	7.4	7.7	6.6	7.3
12	13.0	21.7	13.0	20.0	7.4	7.7	6.7	7.3
13	12.5	21.8	13.1	20.3	7.4	7.7	6.8	7.2
14	12.6	21.7	13.1	20.2	7.4	7.7	6.8	7.4
15	12.8	21.6	13.0	20.2	7.4	7.7	6.9	7.2
16	12.8	21.9	13.1	20.5	7.4	7.7	6.9	7.3
17	12.8	22.0	13.1	20.4	7.4	7.7	6.9	7.4
18	13.0	21.8	13.1	20.4	7.4	7.7	6.9	7.4
19	13.0	21.6	13.1	20.5	7.4	7.7	6.9	7.4

Table 6.22 Apparent error rates and success rates of the imputations made by RIMCA in sI for simulated data with a low correlation structure

RIMCA in SI Low correlated data	Average apparent error rate	Average apparent success rate	Average overall apparent success rate
MAR HR	**12.82 %** of 16%	**3.18 %** of 16%	19.90 %
MCAR HR	**21.56 %** of 30%	**8.44 %** of 30%	28.12 %
MAR HNR	**13.08 %** of 16%	**2.92 %** of 16%	18.26 %
MCAR HNR	**19.96 %** of 30%	**10.04 %** of 30%	33.47 %
MAR LR	**7.4 %** of 8%	**0.60 %** of 8%	7.50 %
MCAR LR	**7.74 %** of 10%	**2.26 %** of 10%	22.58 %
MAR LNR	**6.61 %** of 8%	**1.39 %** of 8%	17.37 %
MCAR LNR	**7.29 %** of 10%	**2.71 %** of 10%	27.05 %

On average, irrespective of the simulation protocol followed per dataset, 21.8% of all the imputations made by RIMCA as a SI procedure are correct for the simulated data with a low correlation structure. The imputations made for the simulated data with MCAR values perform better than the MAR cases when taking the same pattern and percentage of missing values into consideration. This was also observed for the MI procedures.

Table 6.23 Apparent error rates: RIMCA in SI with a high correlation structure

# Dimensions	High percentage of missing values				Low percentage of missing values			
	Random		**Non-random**		**Random**		**Non-random**	
	MAR	MCAR	MAR	MCAR	MAR	MCAR	MAR	MCAR
	16%	30%	16%	30%	8%	10%	8%	10%
1	15.5	20.6	15.4	20.0	7.8	7.7	7.7	7.0
2	15.5	19.6	15.4	19.9	7.8	7.7	7.7	7.0
3	15.5	19.5	15.3	19.9	7.8	7.5	7.7	6.9
4	15.5	19.6	15.4	19.9	7.8	7.5	7.7	6.9
5	15.5	19.9	15.2	20.0	7.8	7.6	7.6	7.0
6	15.5	19.7	15.2	20.1	7.8	7.6	7.7	7.0
7	15.5	19.8	15.3	20.1	7.8	7.7	7.7	7.0
8	15.5	19.9	15.3	20.1	7.8	7.7	7.7	6.9
9	15.5	19.9	15.3	20.1	7.8	7.7	7.7	6.9
10	15.5	20.0	15.4	20.3	7.8	7.7	7.7	7.0
11	15.5	19.8	15.3	20.3	7.8	7.7	7.8	7.0
12	15.5	19.7	15.3	20.1	7.8	7.7	7.8	7.0
13	15.5	19.7	15.3	20.2	7.8	7.8	7.8	7.0
14	15.5	19.8	15.4	20.3	7.8	7.8	7.7	7.1
15	15.5	19.8	15.3	20.5	7.8	7.8	7.7	7.1
16	15.5	19.9	15.2	20.3	7.8	7.8	7.7	7.1
17	15.5	20.0	15.3	21.0	7.8	7.8	7.7	7.1
18	15.5	19.9	15.3	20.7	7.8	7.8	7.7	7.1
19	15.5	19.9	15.3	20.9	7.8	7.8	7.7	7.1

Table 6.24 Apparent error rates and success rates of the imputations made by RIMCA in SI for simulated data with a high correlation structure

RIMCA in SI High correlated data	**Average apparent error rate**	**Average apparent success rate**	**Average overall apparent success rate**
MAR HR	**15.5 %** of 16%	**0.50 %** of 16%	3.13 %
MCAR HR	**19.84 %** of 30%	**10.16 %** of 30%	33.86 %
MAR HNR	**15.31 %** of 16%	**0.69 %** of 16%	4.31 %
MCAR HNR	**20.25 %** of 30%	**9.75 %** of 30%	32.51 %
MAR LR	**7.8 %** of 8%	**0.20 %** of 8%	2.50 %
MCAR LR	**7.71 %** of 10%	**2.29 %** of 10%	22.95 %
MAR LNR	**7.71 %** of 8%	**0.29 %** of 8%	3.62 %
MCAR LNR	**7.01 %** of 10%	**2.99 %** of 10%	29.89 %

On average, irrespective of the simulation protocol followed per dataset, 16.6% of all the imputations made by RIMCA as a SI procedure are correct for the simulated data with a high correlation structure. Once again, the imputations made for the simulated data with MCAR values perform better than the MAR cases when taking the same pattern and percentage of missing values into consideration.

Discussion

The error rates of the different dimensions achieved for the same datasets obtained from SI are highly correlated and in some cases equal. It was expected that the smaller dimensions would result in a higher error rate corresponding to underfitting (fuzziness), as well as high error rates for the larger dimensions where possible overfitting could occur. This however is not the case for all of the dimensions and is therefore only based on expectations and prior beliefs.

Since the error rates obtained for the different dimensions are so similar, it again confirms that a randomly selected dimension to be retained for SI is sufficient (**cf**. 6.3). The similarities between the datasets obtained from the different dimensions could be an indication of the success of the regularisation algorithm, which maybe implies that the regularisation term of the algorithm performs adequately over all dimensions in decreasing the variance.

The statement made by Josse *et al.* (2012:114) regarding the performance of RIMCA in SI is contradicted, since according to above mentioned authors RIMCA is expected to perform better under a MAR mechanism, especially when the data has a strong correlation structure. In both the low and high correlation cases, the overall success rates obtained from the MCAR mechanisms are higher, specifically for the high correlated data structure.

6.6.3 Objective two: Conclusion

The percentage of correct imputations varied from approximately 30% for the MI application of RIMCA and approximately 20% for the application of RIMCA in SI. Therefore, MI evidently outperforms SI, since approximately 10% more

data entries are predicted and imputed correctly. The low percentage of correct imputations might be the result of a reserved structure. It will be interesting to investigate the change in the profiles after imputation, in order to determine whether the missing variable values per individual are replaced with the variable values of other individuals in the survey with similar observed results. In other words, two individuals may have had different responses in the complete data, but then after missingness is applied, they are identical. Through imputation, the first of the two might have obtained the second's original values through imputation, while the second might have obtained the first's original values. This would create apparent errors where, essentially, there are none (the two individuals have just been swapped in the data set after imputation). This suggestion is illustrated in the hypothetical example provided in table 6.25–6.27.

Table 6.25 Hypothetical example: observed survey data

Observed data	Var1	Var2	Var3
Individual 1	2	2	1
Individual 2	2	2	3

Table 6.26 Hypothetical example: entered missing values

Incomplete data	Var1	Var2	Var3
Individual 1	2	2	NaN
Individual 2	2	2	NaN

Table 6.27 Hypothetical example: imputed data

Imputed data	Var1	Var2	Var3
Individual 1	2	2	3
Individual 2	2	2	1

In conclusion, this paper strongly advises the use of the RIMCA algorithm in MI, since the procedure performs well in the prediction of missing values. The applicability of the RIMCA algorithm to be used as a multiple imputation model is established and encouraged.

6.7 Simulation summary

The bias, mean square error (MSE) and coverage obtained over a 1000 simulations are summarised with regard to SI, complete-case analysis (CC) and MI. The coverage refers to the percentage of 95% confidence intervals that contain the complete data (CD) estimate, also known as the true mean of the complete data. Again, each simulation consists of a newly generated dataset, then inserted with missing values and imputed or handled with a specific technique. In the case of MI, a set of ten randomly selected dimensions are used to impute the datasets and then combined using Rubin's rules. The SI makes use of a predetermined dimension, which is chosen as ten (**cf**. 5.5 & 6.3).

The results obtained for the MAR mechanisms are displayed in table 6.28–6.35 (**cf**. 6.7.1) and the results obtained for the MCAR mechanisms will follow in table 6.36–6.43 (**cf**. 6.7.2).

6.7.1 MAR mechanisms

Table 6.28 MAR HNR High correlation: summary over 1000 simulations

MAR HNR High correlation	Bias			MSE			Coverage (%)		
	SI	CC	MI	SI	CC	MI	SI	CC	MI
Variable 1	0	0	0	0	0	0	100	100	100
Variable 2	0.3333	0.1759	0.1756	0.1111	0.0309	0.0308	0.4	43.4	37
Variable 3	0.3330	0.1752	0.1749	0.1109	0.0307	0.0306	0.4	43.4	36.7
Variable 4	0.3332	0.1757	0.1753	0.1110	0.0309	0.0307	0.4	45	39
Variable 5	0.3325	0.1755	0.1751	0.1105	0.0308	0.0306	0.9	45.6	38.8
Variable 6	0.3338	0.1757	0.1752	0.1114	0.0309	0.0307	0.3	44	36.6
Variable 7	0	0	0	0	0	0	100	100	100
Variable 8	0.3129	0.1680	0.1677	0.0979	0.0282	0.0281	2.8	60.8	49.9
Variable 9	0.3128	0.1685	0.1680	0.0979	0.0284	0.0282	3.5	58.4	49.7
Variable 10	0.3110	0.1684	0.1679	0.0967	0.0284	0.0282	4.6	62.7	53.3

Table 6.29 MAR HNR Low correlation: summary over 1000 simulations

MAR HNR Low correlation	Bias			MSE			Coverage (%)		
	SI	CC	MI	SI	CC	MI	SI	CC	MI
Variable 1	0	0	0	0	0	0	100	100	100
Variable 2	0.2298	0.0823	0.0820	0.0520	0.0067	0.0066	16.3	99.5	99.2
Variable 3	0.2316	0.0828	0.0828	0.0526	0.0068	0.0068	16	99.7	99.6
Variable 4	0.2247	0.0809	0.0807	0.0496	0.0064	0.0064	18.6	99.7	99.2
Variable 5	0.2294	0.0816	0.0813	0.0518	0.0066	0.0065	15.7	99.6	99.2
Variable 6	0.2286	0.0825	0.0823	0.0520	0.0067	0.0067	16.4	98.9	98.5
Variable 7	0	0	0	0	0	0	100	100	100
Variable 8	0.1935	0.0784	0.0782	0.0317	0.0060	0.0060	35.7	99.9	99.6
Variable 9	0.1910	0.0780	0.0777	0.0308	0.0060	0.0059	37.1	99.9	99.6
Variable 10	0.1951	0.0780	0.0776	0.0333	0.0060	0.0059	35.7	99.8	99.5

Table 6.30 MAR HR High correlation: summary over 1000 simulations

MAR HR High correlation	Bias			MSE			Coverage (%)		
	SI	CC	MI	SI	CC	MI	SI	CC	MI
Variable 1	0	0	0	0	0	0	100	100	100
Variable 2	0.3326	0.1742	0.1741	0.1106	0.0303	0.0303	0.5	49.3	42.5
Variable 3	0.3338	0.1746	0.1742	0.1114	0.0305	0.0304	0.4	48	42.2
Variable 4	0.3327	0.1752	0.1751	0.1107	0.0307	0.0307	0.5	46.8	39.9
Variable 5	0.3336	0.1746	0.1742	0.1113	0.0305	0.0303	0.4	47.3	41.2
Variable 6	0.3339	0.1759	0.1753	0.1115	0.0309	0.0307	0.7	44.4	38.2
Variable 7	0	0	0	0	0	0	100	100	100
Variable 8	0.3269	0.1749	0.1746	0.1069	0.0306	0.0305	3.4	52.4	42.2
Variable 9	0.3231	0.1754	0.1750	0.1044	0.0307	0.0306	3.5	53.2	45
Variable 10	0.3261	0.1750	0.1746	0.1063	0.0306	0.0305	2.7	52	42.4

Table 6.31 MAR HR Low correlation: summary over 1000 simulations

MAR HR Low correlation	Bias			MSE			Coverage (%)		
	SI	CC	MI	SI	CC	MI	SI	CC	MI
Variable 1	0	0	0	0	0	0	100	100	100
Variable 2	0.2361	0.0843	0.0843	0.0550	0.0070	0.0070	13.8	99.5	98.8
Variable 3	0.2355	0.0828	0.0825	0.0554	0.0068	0.0068	13.3	99.6	98.8
Variable 4	0.2302	0.0846	0.0845	0.0522	0.0070	0.0070	17.5	99.6	98.4
Variable 5	0.2306	0.0838	0.0836	0.0526	0.0069	0.0069	16.3	99.6	99.1
Variable 6	0.2348	0.0856	0.0852	0.0545	0.0072	0.0072	14.8	99.4	99.1
Variable 7	0	0	0	0	0	0	100	100	100
Variable 8	0.2080	0.0833	0.0830	0.0392	0.0069	0.0068	33.1	99.4	98.4
Variable 9	0.2067	0.0830	0.0825	0.0368	0.0068	0.0067	33.1	99.2	98.1
Variable 10	0.1985	0.0822	0.0821	0.0348	0.0067	0.0067	36.5	99.3	99

Table 6.32 MAR LNR High correlation: summary over 1000 simulations

MAR LNR High correlation	Bias			MSE			Coverage (%)		
	SI	CC	MI	SI	CC	MI	SI	CC	MI
Variable 1	0	0	0	0	0	0	100	100	100
Variable 2	0.1602	0.0774	0.0774	0.0257	0.0060	0.0060	46.1	100	100
Variable 3	0.1615	0.0780	0.0777	0.0261	0.0061	0.0060	47.7	100	100
Variable 4	0.1606	0.0780	0.0777	0.0258	0.0061	0.0060	46.1	100	100
Variable 5	0.1596	0.0776	0.0774	0.0255	0.0060	0.0060	48.1	100	100
Variable 6	0.1618	0.0786	0.0783	0.0262	0.0062	0.0061	43.4	100	100
Variable 7	0	0	0	0	0	0	100	100	100
Variable 8	0.1191	0.0738	0.0736	0.0137	0.0055	0.0054	77.1	100	100
Variable 9	0.1179	0.0745	0.0740	0.0137	0.0055	0.0055	80	100	100
Variable 10	0.1192	0.0742	0.0738	0.0140	0.0055	0.0054	78.9	100	100

Table 6.33 MAR LNR Low correlation: summary over 1000 simulations

MAR LNR Low correlation	Bias			MSE			Coverage (%)		
	SI	CC	MI	SI	CC	MI	SI	CC	MI
Variable 1	0	0	0	0	0	0	100	100	100
Variable 2	0.1073	0.0380	0.0381	0.0106	0.0013	0.0013	95.4	100	100
Variable 3	0.1119	0.0400	0.0399	0.0120	0.0015	0.0015	94.1	100	100
Variable 4	0.1054	0.0382	0.0382	0.0106	0.0013	0.0013	95	100	100
Variable 5	0.1063	0.0392	0.0392	0.0105	0.0014	0.0014	95.4	100	100
Variable 6	0.1064	0.0381	0.0380	0.0107	0.0013	0.0013	96.2	100	100
Variable 7	0	0	0	0	0	0	100	100	100
Variable 8	0.0759	0.0373	0.0374	0.0023	0.0012	0.0012	96.1	100	100
Variable 9	0.0755	0.0378	0.0375	0.0024	0.0013	0.0012	96.4	100	100
Variable 10	0.0740	0.0368	0.0368	0.0021	0.0012	0.0012	97.6	100	100

Table 6.34 MAR LR High correlation: summary over 1000 simulations

MAR LR High correlation	Bias			MSE			Coverage (%)		
	SI	CC	MI	SI	CC	MI	SI	CC	MI
Variable 1	0	0	0	0	0	0	100	100	100
Variable 2	0.1652	0.0804	0.0801	0.0273	0.0065	0.0064	40.9	100	100
Variable 3	0.1644	0.0808	0.0808	0.0270	0.0065	0.0065	41.5	100	100
Variable 4	0.1648	0.0809	0.0804	0.0271	0.0065	0.0065	39.5	100	100
Variable 5	0.1641	0.0806	0.0804	0.0269	0.0065	0.0065	41.4	100	100
Variable 6	0.1645	0.0805	0.0804	0.0271	0.0065	0.0065	41	100	100
Variable 7	0	0	0	0	0	0	100	100	100
Variable 8	0.1288	0.0780	0.0776	0.0165	0.0061	0.0060	68.7	100	100
Variable 9	0.1270	0.0783	0.0781	0.0160	0.0061	0.0061	70.7	100	100
Variable 10	0.1244	0.0782	0.0780	0.0154	0.0061	0.0061	73	100	100

Table 6.35 MAR LR Low correlation: summary over 1000 simulations

MAR LR Low correlation	Bias			MSE			Coverage (%)		
	SI	CC	MI	SI	CC	MI	SI	CC	MI
Variable 1	0	0	0	0	0	0	100	100	100
Variable 2	0.1073	0.0388	0.0388	0.0109	0.0014	0.0013	91.5	100	100
Variable 3	0.1086	0.0405	0.0402	0.0109	0.0015	0.0014	91.8	100	100
Variable 4	0.1059	0.0386	0.0384	0.0108	0.0014	0.0013	91	100	100
Variable 5	0.1094	0.0391	0.0391	0.0112	0.0014	0.0014	92	100	100
Variable 6	0.1063	0.0385	0.0384	0.0108	0.0014	0.0013	92.7	100	100
Variable 7	0	0	0	0	0	0	100	100	100
Variable 8	0.0788	0.0377	0.0377	0.0024	0.0013	0.0012	93.7	100	100
Variable 9	0.0820	0.0390	0.0388	0.0029	0.0014	0.0013	92.6	100	100
Variable 10	0.0801	0.0389	0.0387	0.0027	0.0013	0.0013	93.5	100	100

Discussion

In the presence of a high percentage of missing values in the data (**cf**. Table 6.28–6.31) the bias and MSE values for all variables are smaller for low correlated data than for highly correlated data. Furthermore, the coverage is greater for data with a low correlation structure. This is the case for SI, CC and MI techniques. When a smaller percentage of missing values occurs in the data, it is observed that a smaller bias, larger MSE and better coverage is obtained from data with a low correlation structure, also for all of the techniques. Therefore the opposite is true for the high correlated data; larger bias, smaller MSE and smaller coverage in comparison to low correlated data.

SI delivers poor coverage especially for a high percentage of missing values and when a non-random pattern of missing values is entered. As the amount of missing values decreases, the coverage of the SI procedure increases. CC provides larger coverage than MI, but in a majority of the cases the bias and MSE's obtained from MI are slightly smaller than those of CC.

In general, it is clear that MI provides more satisfactory results than SI, with smaller bias and MSE and larger coverage.

6.7.2 MCAR mechanisms

Table 6.36 MCAR HNR High correlation: summary over 1000 simulations

MCAR HNR High correlation	Bias			MSE			Coverage (%)		
	SI	CC	MI	SI	CC	MI	SI	CC	MI
Variable 1	0.3899	0.0822	0.0823	0.0036	0.0000	0.0000	35.7	99.1	94.9
Variable 2	0.3896	0.0818	0.0817	0.0036	0.0000	0.0000	35.4	98.3	94
Variable 3	0.3877	0.0803	0.0805	0.0025	0.0000	0.0000	36.8	98.6	94
Variable 4	0	0	0	0	0	0	100	100	100
Variable 5	0	0	0	0	0	0	100	100	100
Variable 6	0	0	0	0	0	0	100	100	100
Variable 7	0	0	0	0	0	0	100	100	100
Variable 8	0	0	0	0	0	0	100	100	100
Variable 9	0.3948	0.0790	0.0787	0.0046	0.0000	0.0000	33.5	98.5	95
Variable 10	0.4011	0.0812	0.0810	0.0035	0.0000	0.0000	33.1	98.5	94.5

Table 6.37 MCAR HNR Low correlation: summary over 1000 simulations

MCAR HNR Low correlation	Bias			MSE			Coverage (%)		
	SI	CC	MI	SI	CC	MI	SI	CC	MI
Variable 1	0.3910	0.0804	0.0804	0.0045	0.0000	0.0000	36.4	99.2	94.9
Variable 2	0.4027	0.0832	0.0835	0.0015	0.0000	0.0000	34.2	98.6	94.2
Variable 3	0.3972	0.0824	0.0825	0.0046	0.0000	0.0000	35.2	98.3	94.2
Variable 4	0	0	0	0	0	0	100	100	100
Variable 5	0	0	0	0	0	0	100	100	100
Variable 6	0	0	0	0	0	0	100	100	100
Variable 7	0	0	0	0	0	0	100	100	100
Variable 8	0	0	0	0	0	0	100	100	100
Variable 9	0.4061	0.0816	0.0817	0.0032	0.0000	0.0000	33.3	98.2	94.5
Variable 10	0.3941	0.0833	0.0839	0.0040	0.0000	0.0000	35.1	98.6	94.2

Table 6.38 MCAR HR High correlation: summary over 1000 simulations

MCAR HR High correlation	Bias			MSE			Coverage (%)		
	SI	CC	MI	SI	CC	MI	SI	CC	MI
Variable 1	0.1910	0.0441	0.0440	0.0032	0.0000	0.0000	46.1	100	100
Variable 2	0.1949	0.0421	0.0419	0.0032	0.0000	0.0000	43.8	100	99.8
Variable 3	0.1954	0.0441	0.0441	0.0027	0.0000	0.0000	43.4	100	100
Variable 4	0.1980	0.0429	0.0432	0.0022	0.0000	0.0000	42.7	100	100
Variable 5	0.1973	0.0431	0.0432	0.0042	0.0000	0.0000	42.9	100	99.8
Variable 6	0.1963	0.0423	0.0423	0.0029	0.0000	0.0000	43.6	99.9	99.9
Variable 7	0.1896	0.0416	0.0413	0.0036	0.0000	0.0000	44.6	100	99.8
Variable 8	0.1924	0.0426	0.0424	0.0022	0.0000	0.0000	43.8	99.9	99.9
Variable 9	0.1937	0.0420	0.0420	0.0014	0.0000	0.0000	43.6	100	99.9
Variable 10	0.1953	0.0420	0.0419	0.0040	0.0000	0.0000	43.2	100	99.9

Table 6.39 MCAR HR Low correlation: summary over 1000 simulations

MCAR HR Low correlation	Bias			MSE			Coverage (%)		
	SI	CC	MI	SI	CC	MI	SI	CC	MI
Variable 1	0.2048	0.0430	0.0434	0.0026	0.0000	0.0000	40.6	99.8	99.8
Variable 2	0.1935	0.0419	0.0419	0.0051	0.0000	0.0000	43.2	100	99.9
Variable 3	0.1941	0.0430	0.0428	0.0047	0.0000	0.0000	43.5	99.9	99.8
Variable 4	0.1972	0.0439	0.0438	0.0018	0.0000	0.0000	42.7	99.9	99.8
Variable 5	0.1944	0.0414	0.0418	0.0028	0.0000	0.0000	43.8	99.9	99.7
Variable 6	0.1901	0.0412	0.0416	0.0024	0.0000	0.0000	45.1	99.9	99.8
Variable 7	0.1997	0.0422	0.0423	0.0046	0.0000	0.0000	42	100	100
Variable 8	0.1931	0.0419	0.0420	0.0037	0.0000	0.0000	43.9	100	100
Variable 9	0.1941	0.0418	0.0420	0.0030	0.0000	0.0000	42.7	100	99.9
Variable 10	0.1935	0.0411	0.0411	0.0044	0.0000	0.0000	44.2	100	100

Table 6.40 MCAR LNR High correlation: summary over 1000 simulations

MCAR LNR High correlation	Bias			MSE			Coverage (%)		
	SI	CC	MI	SI	CC	MI	SI	CC	MI
Variable 1	0.1331	0.0330	0.0328	0.0027	0.0000	0.0000	52.4	100	100
Variable 2	0.1286	0.0330	0.0332	0.0025	0.0000	0.0000	53.9	100	100
Variable 3	0.1368	0.0323	0.0326	0.0030	0.0000	0.0000	49.2	100	100
Variable 4	0	0	0	0	0	0	100	100	100
Variable 5	0	0	0	0	0	0	100	100	100
Variable 6	0	0	0	0	0	0	100	100	100
Variable 7	0	0	0	0	0	0	100	100	100
Variable 8	0	0	0	0	0	0	100	100	100
Variable 9	0.1314	0.0334	0.0335	0.0018	0.0000	0.0000	53	100	100
Variable 10	0.1285	0.0340	0.0340	0.0016	0.0000	0.0000	54.6	100	100

Table 6.41 MCAR LNR Low correlation: summary over 1000 simulations

MCAR LNR Low correlation	Bias			MSE			Coverage (%)		
	SI	CC	MI	SI	CC	MI	SI	CC	MI
Variable 1	0.1347	0.0320	0.0322	0.0028	0.0000	0.0000	50.7	100	100
Variable 2	0.1361	0.0328	0.0330	0.0029	0.0000	0.0000	49	100	100
Variable 3	0.1377	0.0331	0.0336	0.0032	0.0000	0.0000	49.5	100	100
Variable 4	0	0	0	0	0	0	100	100	100
Variable 5	0	0	0	0	0	0	100	100	100
Variable 6	0	0	0	0	0	0	100	100	100
Variable 7	0	0	0	0	0	0	100	100	100
Variable 8	0	0	0	0	0	0	100	100	100
Variable 9	0.1334	0.0323	0.0326	0.0024	0.0000	0.0000	51.4	99.9	99.9
Variable 10	0.1305	0.0330	0.0329	0.0018	0.0000	0.0000	52.2	100	100

Table 6.42 MCAR LR High correlation: summary over 1000 simulations

MCAR LR High correlation	Bias			MSE			Coverage (%)		
	SI	CC	MI	SI	CC	MI	SI	CC	MI
Variable 1	0.0637	0.0220	0.0222	0.0010	0.0000	0.0000	95.1	100	100
Variable 2	0.0624	0.0204	0.0208	0.0010	0.0000	0.0000	96	100	100
Variable 3	0.0677	0.0225	0.0225	0.0010	0.0000	0.0000	94.6	100	100
Variable 4	0.0631	0.0217	0.0216	0.0008	0.0000	0.0000	96	100	100
Variable 5	0.0664	0.0210	0.0210	0.0013	0.0000	0.0000	96.1	100	100
Variable 6	0.0637	0.0215	0.0216	0.0011	0.0000	0.0000	96.3	100	100
Variable 7	0.0683	0.0218	0.0219	0.0014	0.0000	0.0000	94.9	100	100
Variable 8	0.0666	0.0218	0.0220	0.0011	0.0000	0.0000	95	100	100
Variable 9	0.0660	0.0218	0.0220	0.0010	0.0000	0.0000	95.4	100	100
Variable 10	0.0658	0.0218	0.0217	0.0011	0.0000	0.0000	94.9	100	100

Table 6.43 MCAR LR Low correlation: summary over 1000 simulations

MCAR LR Low correlation	Bias			MSE			Coverage (%)		
	SI	CC	MI	SI	CC	MI	SI	CC	MI
Variable 1	0.0659	0.0208	0.0208	0.0010	0.0000	0.0000	94.9	100	100
Variable 2	0.0678	0.0220	0.0220	0.0011	0.0000	0.0000	94.9	100	100
Variable 3	0.0680	0.0216	0.0217	0.0009	0.0000	0.0000	95.1	100	100
Variable 4	0.0672	0.0216	0.0219	0.0010	0.0000	0.0000	95.7	100	100
Variable 5	0.0666	0.0220	0.0221	0.0010	0.0000	0.0000	95.1	100	100
Variable 6	0.0673	0.0221	0.0223	0.0011	0.0000	0.0000	94.9	100	100
Variable 7	0.0661	0.0217	0.0219	0.0013	0.0000	0.0000	95.7	100	100
Variable 8	0.0668	0.0221	0.0223	0.0009	0.0000	0.0000	95.1	100	100
Variable 9	0.0684	0.0217	0.0219	0.0009	0.0000	0.0000	94.8	100	100
Variable 10	0.0671	0.0214	0.0214	0.0012	0.0000	0.0000	95	100	100

Discussion

There is not a significant difference between the low and high correlation simulations with respect to each specific type of simulated dataset, as was the case for the MAR mechanisms. The difference between bias, MSE and coverage between correlation structures are only slight, therefore no conclusion can be drawn as to which correlation structure provides more satisfying results. It is observed from table 6.36–6.43 that the bias and MSE for SI are larger than the bias and MSE values obtained from MI. The coverage provided by MI is nearly perfect, with an exception of the MCAR HNR datasets, where a slightly smaller

percentage of coverage between 94% and 95% are obtained. Again, the CC measures closely resemble the MI measures.

6.7.3 Simulation summary: Conclusion

In all of the cases of simulated data it is observed that the coverage provided by MI is larger than SI. It is also observed that the bias and MSE values for MI are smaller than the bias and MSE values obtained from SI, this was the case for all variables in all datasets. It seems as if RIMCA in general performed better under a MCAR mechanism (**cf**. Table 6.28–6.31), since smaller MSE values are obtained and the coverage is slightly larger for the CC and MI cases and significantly larger for the SI cases, in comparison to the MAR results (**cf**. table 6.36–6.43).

Therefore, the information provided in sections 6.7.3 and 6.7.4 confirms the success of RIMCA in MI over RIMCA in SI.

6.8 Conclusion

This concludes the simulation study chapter. A motivation for the choice of simulated data was provided, followed by the selection of the specific dimensions used for the reconstruction step of the RIMCA algorithm. The results obtained for objective one and two applied to simulated data were provided and discussed. It was found that RIMCA in MI provides larger apparent success rates than RIMCA in SI; therefore predictions made by MI are more accurate than those obtained by SI. Further, RIMCA in MI produced wider confidence intervals compared to RIMCA in SI, which shows the incorporation of uncertainty. Furthermore, the mean estimates obtained from RIMCA in MI where closer to the true mean in most of the cases. This was followed by a summary of a 1000 simulations, which provided additional information for the success of RIMCA in MI over RIMCA in SI. It was found that larger coverage and smaller bias and mean square errors were obtained from the MI estimates. In the following chapter the real data analysis will be provided and discussed.

Chapter 7
Real Categorical Dataset
Canal des Deux Mers

"The model must fit the data, not vice versa" – Paul-Jean Benzécri

(Greenacre 1984:10)

7.1 Introduction

This chapter will provide the motivation for making use of the specific real dataset. The selected dimensions for the reconstruction step of the RIMCA algorithm to generate imputed datasets will be discussed and motivated. This will be followed by the results obtained from the execution of objective one; concluding with the discussion of the real data results.

7.2 Motivation

The use of the *Canal des Deux Mers* dataset enables the researcher to compare the performance of the RIMCA in MI against the published results (Josse *et al.* 2012) of RIMCA in SI. This allows the researcher to determine which imputation approach (SI or MI) performs better with respect to the RIMCA algorithm. The dataset is obtained from the R package *missMDA* (Husson & Josse 2013).

7.3 Dimensions to retain in the second step of RIMCA

SI

The dimension selected *a priori* for the reconstruction step of the algorithms in SI is based upon the expectance of underfitting (fuzziness) in the lower dimensions and possible overfitting in the larger dimensions, as mentioned by Josse *et al.* (2012:101). Therefore an average number of dimensions are selected, which in the case of the real data is ten dimensions.

MI

As was found in the simulation study (**cf**. 6.3), RIMCA in MI provides stable results across a selection of randomly chosen dimensions. Therefore the random selection of dimensions is carried out for the real data as well. The set of multiple datasets used for further analysis is obtained by randomly selecting ten possible dimensions by means of random integer generation function in MATLAB *(randi)*. The dimensions obtained for RIMCA in MI are 1, 2, 3, 8, 9, 10, 13, 17, 18 and 19 for the real data. The dimensions obtained for IMCA in MI are 2, 3, 4, 5, 8, 9, 10, 13, 15, 21.

7.4 IMCA vs. RIMCA in MI

Table 7.1 Confidence interval widths, means and standard errors obtained from complete-case analysis, IMCA in MI and RIMCA in MI

Real data	Confidence Interval Width			Mean		
Variable	CC	IMCA	RIMCA	CC	IMCA	RIMCA
1	0.0738	0.0736	0.0734	1.4300	1.4296	1.4301
2	0.0988	0.0984	0.0980	1.9085	1.9864	1.9839
3	0.0641	0.0641	0.0637	1.2928	1.2928	1.2926
4	0.0500	0.0501	0.0502	1.2676	1.2676	1.2675
5	0.0495	0.0493	0.0494	1.2575	1.2573	1.2577
6	0.0416	0.0414	0.0414	1.1610	1.1609	1.1611
7	0.0567	0.0569	0.0565	1.4741	1.4743	1.4746
8	0.0567	0.0567	0.0562	1.4095	1.4087	1.4082
9	0.0764	0.0761	0.0754	1.9383	1.9392	1.9390
10	0.0626	0.0620	0.0606	1.4017	1.3993	1.4033
11	0.0640	0.0615	0.0622	1.4491	1.4501	1.4506
12	0.0785	0.0761	0.0731	2.0155	2.0150	2.0157
13	0.0827	0.0828	0.0827	2.1938	2.1943	2.1937
14	0.0635	0.0632	0.0632	1.7150	1.7150	1.7153

Real data	Standard Errors		
Variable	CC	IMCA	RIMCA
1	0.0188	0.0188	0.0187
2	0.0252	0.0251	0.0250
3	0.0163	0.0164	0.0163
4	0.0127	0.0128	0.0128
5	0.0126	0.0126	0.0126
6	0.0106	0.0106	0.0106
7	0.0144	0.0145	0.0144
8	0.0144	0.0145	0.0143
9	0.0195	0.0194	0.0192
10	0.0159	0.0158	0.0154
11	0.0163	0.0157	0.0158
12	0.0200	0.0194	0.0186
13	0.0211	0.0211	0.0211
14	0.0162	0.0161	0.0161

CC – Complete-case analysis

In figure 7.1 the estimated mean of the incomplete real data will be referred to as *CC Mean.* The estimated means will be indicated by *RIMCA Mean* and *IMCA Mean*, indicating the RIMCA estimate and the IMCA estimate, respectively.

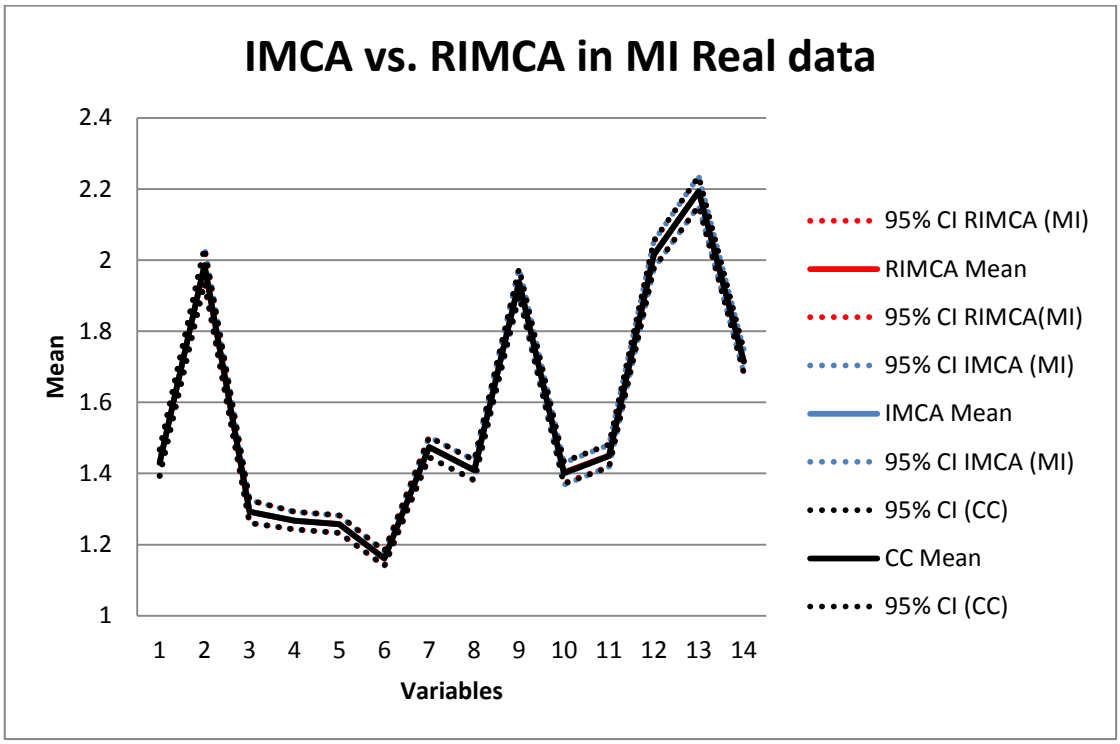


Figure 7.1 Means and Confidence intervals for IMCA and RIMCA in MI

Figure 7.1 illustrates the relationship between the means and confidence intervals obtained from IMCA and RIMCA in MI. As was predicted (Josse *et al.* 2012:103) the results from IMCA and RIMCA will be closely correlated when only a small percentage of missing values occur (approximately 9% missing values) and when a strong correlation structure in the data is present. This shows that overfitting is not a concern and both algorithms will perform well when data is highly correlated and a majority of observed data entries occur. The goal of this dissertation is however to determine the performance of RIMCA in MI compared to RIMCA in SI. This graph is added for interest's sake and to confirm the results obtained by Josse *et al*. (2012).

7.5 Objective one: To establish whether RIMCA in MI outperforms RIMCA in SI

7.5.1 RIMCA in MI vs. SI

The comparison between the results obtained from RIMCA in MI and SI will be given by means of table 7.2 and figure 7.2. Again, the estimated mean of the incomplete real data will be referred to as *CC Mean*. The estimated means obtained from the imputation procedures are indicated by *MI Mean* and *SI Mean*.

Table 7.2 Confidence interval widths, means and standard errors obtained from complete-case analysis, RIMCA in SI and RIMCA in MI

Real data	Confidence Interval Width			Mean		
Variable	CC	SI	MI	CC	SI	MI
1*	0.0738	0.0714	**0.0734**	1.4303	1.4131	1.4301
2*	0.0988	0.0961	**0.0980**	1.9852	1.9156	1.9839
3*	0.0641	0.0623	**0.0637**	1.2928	1.2833	1.2926
4*	0.0500	0.0492	**0.0502**	1.2676	1.2622	1.2675
5*	0.0495	0.0485	**0.0494**	1.2575	1.2516	1.2577
6*	0.0416	0.0407	**0.0414**	1.1610	1.1575	1.1611
7*	0.0567	0.0557	**0.0565**	1.4741	1.4602	1.4746
8*	0.0567	0.0544	**0.0562**	1.4095	1.3856	1.4082
9*	0.0764	0.0674	**0.0754**	1.9383	1.9456	1.9390
10*	0.0626	0.0517	**0.0606**	1.4017	1.3084	1.4033
11*	0.0640	0.0530	**0.0622**	1.4491	1.3401	1.4506
12*	0.0785	0.0534	**0.0731**	2.0155	2.0106	2.0157
13*	0.0827	0.0808	**0.0827**	2.1938	2.1891	2.1937
14	0.0635	0.0714	0.0632	1.7150	1.4131	1.7153

Real data	Standard Errors		
Variable	CC	SI	MI
1	0.0188	0.0182	0.0187
2	0.0252	0.0245	0.0250
3	0.0163	0.0159	0.0163
4	0.0127	0.0125	0.0128
5	0.0126	0.0124	0.0126
6	0.0106	0.0104	0.0106
7	0.0144	0.0142	0.0144
8	0.0144	0.0139	0.0143
9	0.0195	0.0172	0.0192
10	0.0159	0.0132	0.0154
11	0.0163	0.0135	0.0158
12	0.0200	0.0136	0.0186
13	0.0211	0.0206	0.0211
14	0.0162	0.0156	0.0161

CC – complete-case analysis, SI – single imputation, MI – multiple imputation
* – indicates the variables with a wider confidence interval with regard to MI (only considering SI and MI)

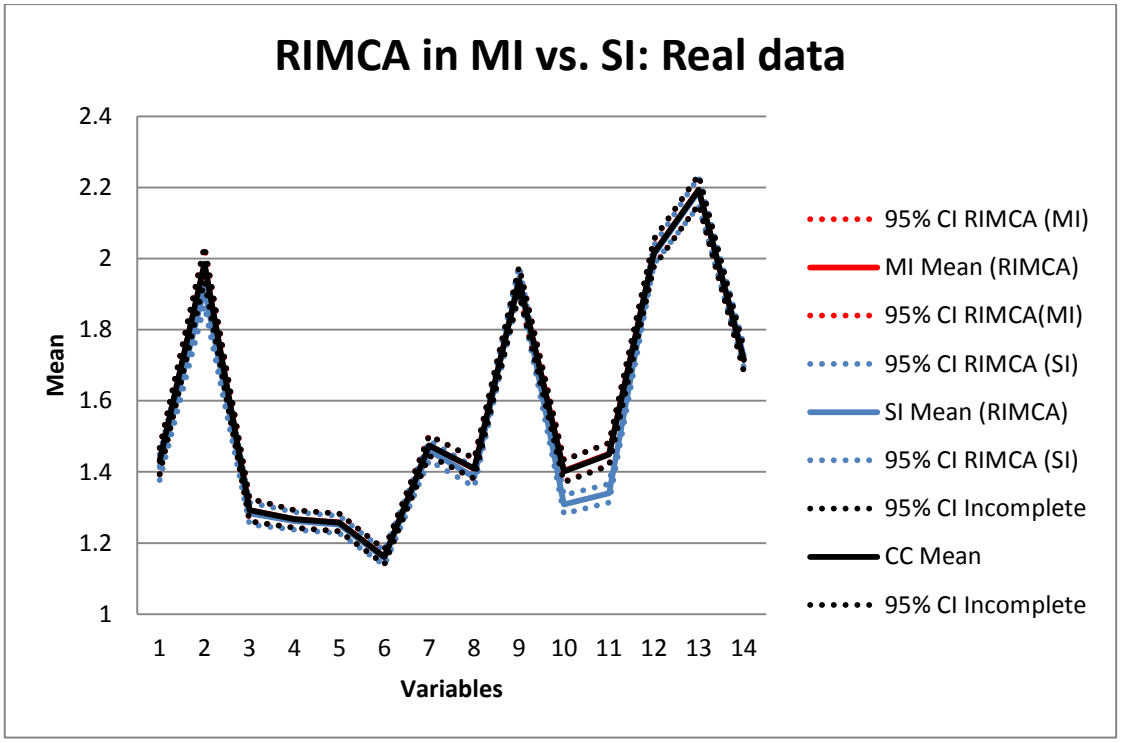


Figure 7.2 Means and Confidence intervals for RIMCA in MI and SI

Figure 7.2 illustrates the comparison between RIMCA in MI and RIMCA in SI. The estimated means are similar with a few slight deviations. The confidence intervals for MI are slightly wider for all of the variables, with the exception of variable 14, where the SI confidence interval is wider. However the uncertainty added by MI is only slight as can be seen from the between variance *(B)* statistics provided in appendix M.1. The between variance indicates that the differences in estimates across the different dimensions are not significant. This confirms the choice of making use of a random selection of dimensions for the generation of the multiple datasets (**cf**. 7.3). The information obtained from the use of Rubin's rules (**cf**. 3.5) is presented in table 7.3. Seven of the 14 variables of the real data will be displayed (**cf**. Appendix M.1).

Table 7.3 Rubin's rules for RIMCA in MI on real data

RIMCA	Variables						
MI	1	2	3	4	5	6	7
Q bar	1.4301	1.9839	1.2926	1.2675	1.2577	1.1611	1.4746
U bar	0.0003	0.0006	0.0003	0.0002	0.0002	0.0001	0.0002
B	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
T	0.0004	0.0006	0.0003	0.0002	0.0002	0.0001	0.0002
df	52548.90	15486.17	75691.62	63364.36	111631.60	160231.13	80662.69
fmi	0.0306	0.0564	0.0255	0.0278	0.0210	0.0175	0.0247
r	0.0315	0.0596	0.0261	0.0286	0.0214	0.0178	0.0253
eff	0.9997	0.9994	0.9997	0.9997	0.9998	0.9998	0.9998
t-val	1.9600	1.9601	1.9600	1.9600	1.9600	1.9600	1.9600
95% CI	1.3935	1.9349	1.2607	1.2425	1.2330	1.1404	1.4463
95% CI	1.4668	2.0328	1.3244	1.2926	1.2824	1.1818	1.5028
width	0.0734	0.0980	0.0637	0.0502	0.0494	0.0414	0.0565

Q bar –estimated mean for MI (Rubin's rules), U bar –estimated variance for MI (Rubin's rules), B – between-imputation variance, T – total variance, df – degrees of freedom, fmi – fraction of missing information, r – relative increase in variance, eff – efficiency rate, 95% CI – 95% confidence intervals, width – confidence interval width

The fmi, r and eff measures show that the amount of information provided by the missing values is small, since the fmi and r measures are small. The r measure represents the relative increase in the variance due to the missing data entries and the fmi measure represents the fraction of information regarding the mean (Q) due to the missing data entries. Furthermore, the efficiency measure (eff) is large (in some cases perfect) across all variables. Schenker, Raghunathan, Chiu, Makuc, Zhang and Cohen (2006:929) mention that SI and MI will produce point estimates which are very closely related when the fraction of missing information is small. This can also be the result of small between-imputation variances (B_m) that occur when little information is missing. From table 7.3 it can be observed that the variance measures U, B_m and T, are very small, close to zero (only two decimal places are indicated). Also, the rate of missing information tends to be less than the percentage of missing data, which can be explained by the correlation between variables and the success of predicting missing values from the observed values (Rubin 1988:82).

RIMCA in MI adds uncertainty, resulting in wider confidence intervals for the mean. The performance of SI and MI is very similar in the case of this specific dataset. The similarity of the two procedures (SI and MI) might be due to the

small number of missing values in the data (9%) and also because of a high correlation between the variables.

It will be interesting to investigate the difference in performance of these methods when applied to data with a large percentage of missing values and where weak correlation is present.

7.6 Conclusion

In this chapter the results obtained from the application of the IMCA and RIMCA algorithm were illustrated and discussed. The dimensions retained for the second step of the algorithms were provided and motivated. A comparison of the performance of IMCA and RIMCA in MI were provided, as well as the results obtained for the completion of objective one, concerning the real data. It was found that the estimates obtained from IMCA and RIMCA in MI were extremely close, which could be explained by a high correlation structure and a low percentage of missing values in the data. Further it was found that RIMCA in MI provides estimates close to the incomplete data estimates, whilst providing wider confidence intervals than SI. In the following and final chapter of this dissertation the discussion of all the chapters and the conclusion of the study will be provided.

Chapter 8
Discussion and Conclusion

We shall not cease from exploration

And the end of all our exploring

Will be to arrive where we started

And know the place for the first time

T.S. Eliot (1943)

8.1 Introduction

This final chapter includes the conclusions and recommendations aimed at further / future research in the field of MI in missing survey data.

Non-responses in data are a pervasive problem, especially experienced in survey data. MI techniques increase the accuracy of estimations, valid inferences are attained when combining the multiple datasets and finally, MI enables imputations that are repeatedly randomly drawn under a number of models. Thus the sensitivity of inferences to a variety of models for missing values can be studied by simply repeatedly using complete-case methods (Rubin 1978:16 & **cf**. 3.4.3.2).

The use of a MCA algorithm as an imputation procedure is appropriate in the context of missing survey data, since MCA is concerned with the similarities and associations within a set of two or more variables. A regularised iterative process with regard to MCA was developed by Josse *et al.* (2012:99) and used as a SI method. Since MI possesses the above-mentioned advantages over SI, it was an intriguing idea to investigate the performance of RIMCA in MI.

The aim of this study, as described in Chapter One, was to investigate the success of RIMCA in MI. This aim was reached by means of executing two objectives: to establish whether RIMCA in MI outperforms RIMCA in SI and to

investigate the accuracy of the predictions made by RIMCA in MI when applied to a simulated dataset.

8.2 Conclusions

The first objective was completed by performing RIMCA as a SI and MI method on the same datasets, real and simulated. The simulation study consisted of 16 different datasets, varying with regard to correlation structure, missingness mechanisms and percentage of missing values in the data. Therefore, the performance of RIMCA in MI was compared to the performance of RIMCA in SI in 17 different constructed datasets (one real dataset and 16 simulated datasets), in order to provide comprehensive analysis.

From the simulation application and real data analysis dedicated to this objective, it was found that in almost all of the cases the confidence intervals provided by MI were wider than those from SI, which confirms the added uncertainty when multiple datasets are imputed. Also, the accuracy of the mean estimates obtained from MI was closer to the true mean values than the estimates provided by SI. This confirmed that the application of RIMCA in MI provided sufficient results in both the real and simulated datasets.

The second objective was completed by performing the RIMCA algorithm on the simulated data, in order to determine the accuracy of the estimations made by the imputation model as well as the success rate of predictions.

The results obtained showed that, on average, MI provided 10% more successful imputed values than SI. The average success rate of MI, with regard to the calculation of apparent error rates over all dimensions, was approximately 30%. Therefore 70% of the missing values were predicted incorrectly. This error was larger than expected, but the paper provides a reasonable possible explanation for this large error. Regardless of the low success rate, MI outperformed SI based on accuracy (**cf**. 6.6.3).

RIMCA in SI experienced two problems (df. 1.2): the uncertainty of which dimensions to choose and that the final fuzzy values obtained from the imputation actually have inherent uncertainties which are not accounted for in

the SI method, both these problems lead to invalid inferences. Since RIMCA in MI imputes multiply over several dimensions, the first problem is overcome. By drawing multiply from the final fuzzy values when allocating categories to the imputed values, several multiple datasets are obtained for one dataset of final fuzzy values; therefore the second problem is also overcome in the MI adaptation of the RIMCA algorithm. The amendments on the SI method resulted in valid confidence intervals and improved efficiency.

Therefore, the aim of the study to investigate the success of RIMCA in MI was achieved. It was found that in both the real and simulated data RIMCA performs better in MI than in SI.

8.3 Limitations of the study

A list of assumptions made in this study are considered as the limitations:

- Only ordinal categorical data was considered.
- It was regarded as sufficient to make use of the means as a result.
- It was assumed that the data followed a Normal distribution in the computation of the confidence intervals. The assumption seemed appropriate due to the large degrees of freedom obtained.
- As for any incomplete dataset, no fully observed (complete) dataset is available for the comparison of imputed data with the original data. Therefore, a limitation experienced in the presence of non-responses is that imputations will always be based on estimates and never truly known.

8.4 Recommendations and further research

- A MNAR mechanism can be added for the simulation study, by deliberately removing certain values, in order to establish the performance of RIMCA in MI with regard to all the missingness mechanisms.

- The difficulty of specifying the missingness mechanism should be addressed for datasets containing a very large number of variables.
- The success of RIMCA in MI with respect to the MAR mechanism can be further confirmed by adding more auxiliary variables from the external sources. This will result in a more plausible assumption of the MAR mechanism.
- RIMCA in MI can be applied to a real dataset with a weaker correlation structure and a higher percentage of missing values.
- Extra bias can be incorporated by a higher percentage of missing values, this might result in a significant difference between the results obtained from complete-case analysis and MI.
- It would be interesting to establish what the role of the initial values is and why they do not contribute to the final imputed values.
- RIMCA in MI can be compared to the multinomial model in sequential regression multiple imputation (SRMI).
- It would be interesting to apply RIMCA as a non-iterative method, therefore a regularised multiple correspondence analysis imputation procedure. This might allow for MNAR values, since the results won't be forced into a MAR mechanism due to the iterations.
- In order to investigate the high apparent error rates for the simulated data, profile analysis can be done for each imputation (**cf**. 6.6.3).
- Other methods may be introduced to analyse ordinal data.
- Methods are required to determine the success of the RIMCA algorithm in nominal categorical data.

8.5 Conclusion

The aim of this study with its objectives was achieved. The use of the regularised iterative multiple correspondence analysis (RIMCA) algorithm in

multiple imputation (MI) provided accurate estimates and wider confidence intervals. Therefore, whilst adding the uncertainty when imputing missing values, sufficient estimates were obtained.

At the conclusion of this dissertation it should be noted that in the context of information being missing, any percentage regained is an accomplishment, since something that was expected to have been lost, has partially been replaced and therefore provides additional information.

"Data forever lost, but never forgotten" – Johané Nienkemper 2013

List of References

Abayomi K, Gelman A & Levy M. 2005. Diagnostics for multivariate imputations. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 57(3):273–291.

Abdi H. 2007. Singular value decomposition (SVD) and generalized singular value decomposition (GSVD). In Neil J. Salking (ed). *Encyclopedia of measurement and statistics,* pp. 907–912. Thousand Oaks, CA: Sage Publications. Available at: http://www.utdallas.edu/~herve/Abdi-SVD2007-pretty.pdf (Accessed 29 May 2013).

Abdi H & Valentin D. 2007. Multiple Correspondence Analysis. In Neil J. Salkind (ed). *Encyclopedia of measurement and statistics,* Vol. 2, pp 425–798. Thousand Oaks, CA: Sage Publication.

Ali MW & Siddiqui O. 2000. Multiple imputation compared with some informative dropout procedures in the estimation and comparison of rates of change in longitudinal clinical trials with dropouts. *Journal of Biopharmaceutical Statistics* 10(2):165–181.

Ambler G, Omar RZ & Royston P. 2007. A comparison of imputation techniques for handling missing predictor values in a risk model with a binary outcome. *Statistical Methods in Medical Research* 16:277–298.

Ardington C, Lam D, Leibbrandt M & Welch M. 2006. The sensitivity to key data imputations of recent estimates of income poverty and inequality in South Africa. *Economic Modelling* 23:822–835.

Blasius J & Greenacre M. 2006. Correspondence analysis and related methods in practice. In Greenacre J & Blasius J (eds). *Multiple Correspondence Analysis and Related Methods.* Taylor & Francis Group, LLC. Boca Raton, London & New York, pp. 4–40.

Blasius J & Thiessen V. 2012. *Assessing the quality of survey data.* SAGE Publications Ltd, London.

Buhi ER, Goodson P & Neilands TB. 2008. Out of Sight, Not Out of Mind: Strategies for Handling Missing Data. *American Journal of Health Behaviour* 32(1):83–92.

Cabras S, Castellanos ME & Quirós A. 2011. Goodness-of-Fit of Conditional Regression Models for Multiple Imputation. *Bayesian Analysis* 6(3):429–456.

Chavent M, Kuentz V & Saracco J. 2010. A partitioning method for the clustering of categorical variables. In Locarek-Junge H & Weihs C (eds). *Classification as a tool for research*. Springer-Verlag. Berlin Heidelberg. Germany, pp. 91–100.

Cormen TH, Leiserson CE, Rivest RL & Stein C. 2001. Introduction to algorithms, 2nd ed. The Massachusetts Institute of Technology. United States of America.

De Tibeiro JJS & Murdoch DJ. 2010. Correspondence Analysis with Incomplete Paired Data using Bayesian Imputation. *Bayesian Analysis* 5(3):519–532.

Eliot TS. 1943. *Four Quartets*. US: Harcourt.

García-Laencina PJ, Figueiras-Vidal AR & Sancho-Gómez J. 2010. Pattern classification with missing data: a review. *Neural Comput & Applic* 19:263–282.

Greenacre M. 2006. From simple to multiple correspondence analysis. In Greenacre J & Blasius J (eds). *Multiple Correspondence Analysis and Related Methods.* Taylor & Francis Group, LLC. Boca Raton, London & New York, pp. 41–76.

Greenacre M. 2007. Correspondence Analysis in Practice, 2nd ed. Chapman & Hall/CRC, Taylor & Francis Group, LLC. Boca Raton.

Greenacre M. 2010. *Biplots in Practice*. Fundación BBVA. Available at http://www.multivariatestatistics.org (Accessed 21 May 2012).

Greenacre MJ. 1984. *Theory and Applications of Correspondence Analysis*. Academic Press, Inc. London.

He Y & Raghunathan TE. 2009. On the Performance of Sequential Regression Multiple Imputation Methods with Non Normal Error Distributions. *Communications in Statistics-Simulation and Computation* 38:856–883.

Heitjan DF & Rubin DB. 1991. Ignorability and coarse data. *The Annals of Statistics* 19(4):2244–2253.

Husson F & Josse J. 2013. missMDA. Handling missing values with/in multivariate analysis (principal component methods). *R package version 1.7.1.* Available at http://CRAN.R-project.org/package=missMDA (Accessed 14 September 2012).

Ientilucci EJ. 2003. Using the Singular Value Decomposition, *Technical Report May, Rochester Institute of Technology*, College of Science, Center for Imaging Science, Digital Imaging and Remote Sensing Laboratory, Rochester, New York, United States. Available at http://astro.rit.edu/~ejipci/Reports/svd.pdf (Accessed 23 January 2013).

Izenman AJ. 2008. *Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning.* Springer Science+Business Media, LLC. New York, United States of America.

Jackson JE. 1991. *A User's Guide To Principal Components*. John Wiley & Sons, Inc. United States of America.

Jamshidian M & Jalal S. 2010. Test of homoscedasticity, normality, and missing completely at random for incomplete multivariate data. *Psychometrika* 75(4):649–674.

Jolliffe IT. 1986. *Principal Component Analysis*. Springer-Verlag. New York, United States of America.

Josse J, Chavent M, Liquet B & Husson F. 2011. Handling missing values with regularized iterative multiple correspondence analysis. *ICC conference notes.* University of St Andrew. United Kingdom. Available at http://math.agrocampus-ouest.fr/infoglueDeliverLive/digitalAssets/47015_josse_husson_liquet_chavent_st_andrews.pdf (Accessed 15 April 2013).

Josse J, Chavent M, Liquet B & Husson F. 2012. Handling missing values with regularized iterative multiple correspondence analysis. *Journal of Classification* 29:91–116.

Josse J & Husson F. 2012. Handling missing values in exploratory multivariate data analysis methods. *Journal de la Société Française de Statistique* 153(3):79–99.

Kenward MG & Carpenter J. 2007. Multiple imputation: current perspectives. *Statistical Methods in Medical Research* 16:199–218.

King G, Honaker J, Joseph A & Scheve K. 2001. Analyzing Incomplete Political Science Data: An Alternative Algorithm for Multiple Imputation. *American Political Science Review* 95(1):49–69.

Klema VC & Laub AJ. 1980. The Singular Value Decomposition: Its Computation and Some Applications. *IEEE Transactions on Automatic Control* 25(2):164–176.

Le Roux B & Rouanet H. 2004. *Geometric Data Analysis: From Correspondence Analysis to Structured Data Analysis.* Springer Science+Business Media, Inc. Kluwer Academic Publishers.

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.

Little RJA & Rubin DB. 2002. *Statistical analysis with missing data*, 2nd ed. Wiley-Interscience, John Wiley & Sons, Inc. Publication. United States of America.

Little R. 2011. Calibrated Bayes, for Statistics in General, and Missing Data in Particular. *Statistical Science* 26(2):162–174.

Lu K, Jiang L & Tsiatis AA. 2010. Multiple Imputation Approaches for the Analysis of Dichotomized Responses in Longitudinal Studies with Missing Data. *Biometrics* 66:1202–1208.

Madsen RE, Hansen LK & Winther O. 2004. Singular value decomposition and principal component analysis. *Technical Report*. Available at http://www2.imm.dtu.dk/pubdb/views/publication_details.php?id=4000 (Accessed 25 July 2012).

Meng X-L. 1994. Multiple-Imputation inferences with uncongenial sources of input. *Statistical Science* 9(4):538–573.

Nenadić O & Greenacre M. 2006. Computation of multiple correspondence analysis, with code in R. In Greenacre J & Blasius J (editors). *Multiple Correspondence Analysis and Related Methods.* Taylor & Francis Group, LLC. Boca Raton, London & New York, pp.523-551.

Penn DA. 2007. Estimating missing values from the general social survey: an application of multiple imputation. *Social Science Quarterly* 88(2):573–584.

Pigott TD. 2001. A review of methods for missing data. *Educational Research and Evaluation* 7(4):353–383.

Quinn G & Keough MJ. 2002. *Experimental design and data analysis for biologists.* Cambridge University Press, United Kingdom.

Raghunathan TE, Lepkowski JM, Van Hoewyk J & Solenberger P. 2001. A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models. *Survey Methodology* 27(1):85–95.

Reiter JP & Raghunathan TE. 2007. The Multiple Adaptations of Multiple Imputation. *Journal of the American Statistical Association* 102(480):1462–1471.

Rencher AC. 2002. *Methods of Multivariate Analysis*, 2nd ed. Wiley-Interscience, John Wiley & Sons, Inc. Publication. United States of America.

Rice JA. 1995. *Mathematical Statistics and Data Analysis*, 2nd ed. Belmont, California: Wadsworth publishing company. United States of America.

Ripley BD. 1987. Stochastic Simulation. John Wiley & sons, Inc. United States of America.

Royston P. 2004. Multiple imputation of missing values. *The Stat Journal* 4(3):227–241.

Rubin DB. 1976. Inference and missing data. *Biometrika* 63(3):581–592.

Rubin DB. 1978. Multiple imputation in sample surveys – a phenomenological Bayesian approach to nonresponse. In *Proceedings of the Survey Research Methods Section*. Washington D.C., pp. 20–34.

Rubin DB. 1987. Multiple imputation for nonresponse in surveys. John Wiley & Sons, Inc. United States of America.

Rubin DB. 1988. An overview of multiple imputation. *Proceedings of the Survey research methods section of the American Statistical Association,* pp. 79–84. Available at http://www.amstat.org/sections/srms/Proceedings/papers/1988_016.pdf (Accessed 14 June 2013)

Rubin DB. 1996. Multiple Imputation After 18+ Years. *Journal of the American Statistical Association* 91(434):473–489.

Rubin DB. 2003a. Discussion on Multiple Imputation. *International Statistical Review* 71(3):619–625.

Rubin DB. 2003b. Nested multiple imputation of NMES via partially incompatible MCMC. *Statistica Neerlandica* 57(1):3–18.

Rubin DB & Schenker N. 1986. Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *Journal of the American Statistical Association* 81(394):366–374.

Schafer JL. 1999. Multiple imputation: a primer. *Statistical Methods in Medical Research* 8:3–15.

Schafer JL. 2003. Multiple imputation in multivariate problems when the imputation and analysis models differ. *Statistica Neerlandica* 57(1):19–35.

Schafer JL & Graham JW. 2002. Missing data: our view of the state of the art. *American Psychological Association, Inc.* 7(2):147–177.

Schafer J & Olsen M. 1998. Multiple Imputation for Multivariate Missing-Data Problems: A Data Analyst's Perspective. *Multivariate Behavioral Research* 33(4):545–571.

Schenker N, Raghunathan TE, Chiu P-L, Makuc DM, Zhang G & Cohen AJ. 2006. Multiple imputation of missing income data in the National Helath Interview survey. *Journal of the American Statistical Association* 101(475):924–933.

Song Q & Shepperd M. 2007. A new imputation method for small software project data sets. *The Journal of Systems and Software* 80:51–62.

Tabachnick BG & Fidell LS. 1989. *Using Multivariate Statistcs,* 2nd ed. HarperCollins*Publisher,* Inc. New York, United State of America.

Takane Y & Hwang H. 2006. Regularized multiple correspondence analysis. In Greenacre J & Blasius J (eds). *Multiple Correspondence Analysis and Related Methods.* Taylor & Francis Group, LLC. Boca Raton, London & New York, pp. 259–279.

Tappen RM. 2011. *Advanced Nursing Research*. Jones & Bartlett Learning, LLC. United States of America.

Van Buuren S, Brand JPL, Groothuis-Oudshoorn CGM & Rubin DB. 2006. Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation* 76(12):1049–1064.

Van der Heijden PGM & Escofier B. 2003. Multiple correspondence analysis with missing data. In Escofier B (ed). *Analyse des correspondances. Recherches au Coeur de I'analyse des donnees.* Rennes: Presses Universitaire de Rennes – Societe Francaise de Statistque, pp. 153–170.

Available at http://www.fss.uu.nl/pubs/pgmvanderheijden/67.multiple correspondenceanlaysis.pdf (Accessed 17 June 2013).

Von Maltitz MJ. 2010. Research Methodology. *Course notes.* University of the Free State.

Von Maltitz MJ & van der Merwe AJ. 2012. An application of sequential regression multiple imputation on panel data. *South African Journal of Economics* 80(1):77–90.

Wall ME, Rechtsteiner A & Rocha LM. 2003. Singular value decomposition and principal component analysis. In Berrar DP, Dubitzky W & Granzow M (eds). *A Practical Approach to Microarray Data Analysis.* Kluwer, Norwell, MA, pp. 91–109.

Wang L & Fan X. 2004. Missing data in disguise and implications for survey data analysis. *Field Methods* 16:332. Available at http://fmx.sagepub.com/content/16/3/332. (Accessed 1 November 2012).

Wayman JC. 2003. Multiple Imputation for missing data: What is it and how can I use it? *Annual Meeting of the American Educational Research Association.* Chicago, IL. Available at http://www.csos.jhu.edu/contact/staff/jwayman_pub/wayman_multimp_aera2003.pdf. (Accessed 31 August 2012).

White IR, Wood A & Royston P. 2007. Multiple imputation in practice. *Statistical Methods in Medical Research* 16:195–197.

Zhang P. 2003. Multiple imputation: theory and method. *International Statistical Review* 71(3):581–592.

Appendices

Appendix A: Functions used within IMCA in RIMCA algorithms

Assigning initial values

```
function [row_vals] = initial_vals(cats,num_missing,varargin)

if nargin > 2 %SI (proportionally)
    randoms = 0;
    data = varargin{1};
else
    randoms = 1;
end

if randoms %MI ( allocate random initial values)
    row_vals=zeros(num_missing,cats);
    for num = 1:num_missing
        row_vals(num,1) = rand;
        for i=2:cats
            if i ~= cats
                row_vals(num,i)=rand*(1-sum(row_vals(num,1:(i-1))));
            else
                row_vals(num,i)=1-sum(row_vals(num,1:(i-1)));
            end
        end
    end
else
row_vals=repmat(nansum(data)./sum(isfinite(data(:,1))),num_missing,1);
end
```

Assigning dummy variables to fuzzy values

```
function [row_vals] = assign01_to_rows(fuzzy_vectors, varargin)
if nargin > 1
    flag_random = 1;%categories allocated randomly (MI)
else
    flag_random = 0;%categories allocated to degree of membership (SI)
end

[num cats] = size(fuzzy_vectors);
row_vals = zeros(num,cats);
if flag_random == 0   %SI
    for i = 1:num
        [~,index] = sort(fuzzy_vectors(i,:),'descend');
        row_vals(i,index(1)) = 1;
    end
else
    for i = 1:num   %MI
        target_vec = cumsum(fuzzy_vectors(i,:));
        target_vec = target_vec./target_vec(end);
        index = find(target_vec > rand, 1, 'first');
        row_vals(i,index(1)) = 1;
    end
end
```

Reconstruction of indicator matrix to data matrix

```
function [data_final] = categorical_recon(X, vec_size);

[I,~] = size(X);
vec_cumsum = cumsum(vec_size);
cols = length(vec_size);
data_final = zeros(I,length(vec_size));

for col = 1:cols
    if col == 1
        Z = X(:,1:vec_size(1));
        for k = 1:vec_size(col)
            indices = find(Z(:,k)==1);
            n = length(indices);
            data_final(indices,col) = k*ones(n,1);
        end
    else
        Z = X(:,(vec_cumsum(col-1)+1):(vec_cumsum(col)));
        for k = 1:vec_size(col)
            indices = find(Z(:,k)==1);
            n = length(indices);
            data_final(indices,col) = k*ones(n,1);
        end
    end
end
```

Appendix B: IMCA algorithm

```
%STEP 1 IMCA
Function
[X_final,X_start,X_hat,place_holders,vec_size]=imca(data,S,threshold)
format long

[I J] = size(data);
X0 = dummyvar(data);
W=1-(isnan(X0));
vec_size=zeros(J,1);
X = X0;

%Assigning initial values
place_holders = cell(1,J);
for col = 1:J
    vec_size(col)=length(unique(data(isfinite(data(:,col)),col)));
    vec_cumsum = cumsum(vec_size);
    place_holders{col} = find(isnan(data(:,col)));
    if col == 1
      %[row_vals]=initial_vals(vec_size(col),length(place_holders{col}
      ),X0(:,1:vec_size(1))); %SI

      [row_vals]=initial_vals(vec_size(col),length(place_holders{col})
      ); %MI
      X(place_holders{col},1:vec_size(1)) = row_vals;
    else
      %[row_vals]=initial_vals(vec_size(col),length(place_holders{col}
      ),X0(:,(vec_cumsum(col-1)+1):(vec_cumsum(col)))); %SI

      [row_vals]=initial_vals(vec_size(col),length(place_holders{col})
      );%MI
```

```
        X(place_holders{col},(vec_cumsum(col-1)+1):(vec_cumsum(col)))
= row_vals;
    end
end

X_start=X;
[~,K]=size(X);
D_sigma=diag(sum(X));
X_hat = X;

%I = # individuals; J = # of variables, K = # of categories
iteration=0;
iterations=100;
change = 1;
conv_prog=(zeros(sum(sum(isnan(data))),max(vec_size),iterations));
conv_prog_plot=zeros(iterations,
max(vec_size),(sum(sum(isnan(data)))));
size(conv_prog);
%Step 3 IMCA (while loop)
while change > threshold
    %Step 2a IMCA
    row_counter=0;
    Z = I*[(X_hat)*(inv(D_sigma))];
    M = (1/(I*J))*D_sigma;
    D=(1/I)*eye(I);

    Z_est=(Z-ones(I,1)*(ones(K,1))')*sqrt(M);
    Z_tilda=D.^(0.5)*Z_est*M.^(0.5);

    [P L Q]=svd(Z_tilda);
    C=D^(-0.5)*P;
    U_hat=M^(-0.5)*Q;
    F_hat=C*L;

    F_hat_dim=F_hat(:,1:S);
    U_hat_dim=U_hat(:,1:S);

    %Step 2b ITERATIVE IMCA
    for i=1:I;
        for k=1:K;
            a=(D_sigma(k,k)/I);
            outer_sum=sum((F_hat_dim(i,(1:S)).*U_hat_dim(k,(1:S))));
            X_hat_iter(i,k) = a*(outer_sum+1);
        end
    end
    change = sum(sum(((1-W).*X_hat - (1-W).*X_hat_iter).^2));
    change_matrix = ((1-W).*X_hat - (1-W).*X_hat_iter).^2;

    X_hat=(W.*X)+((1-W).*X_hat_iter);
    D_sigma=diag(sum(X_hat));
    iteration=iteration+1;

    for col = 1:J
        if col == 1;
            fuzzy_vectors = X_hat(place_holders{col},1:vec_size(1));

change_vectors=change_matrix(place_holders{col},1:vec_size(1));

        else
```

```
            fuzzy_vectors = X_hat(place_holders{col},(vec_cumsum(col-
1)+1):(vec_cumsum(col)));
            change_vectors =
change_matrix(place_holders{col},(vec_cumsum(col-
1)+1):(vec_cumsum(col)));
        end
        a=row_counter+1;
        b=row_counter+length(change_vectors(:,1));
        conv_prog(a:b,1:vec_size(col),iteration)=change_vectors;
        row_counter=b;
        conv_prog_plot=permute(conv_prog,[3 2 1]);
    end
end

X_final = X_hat;
for col=1:J
    if col == 1;
        fuzzy_vectors = X_hat(place_holders{col},1:vec_size(1));
    else
        fuzzy_vectors = X_hat(place_holders{col},(vec_cumsum(col-
1)+1):(vec_cumsum(col)));
    end

    %[row_vals] = assign01_to_rows(fuzzy_vectors); %SI
    [row_vals] = assign01_to_rows(fuzzy_vectors,1); %MI
    if col == 1;
        X_final(place_holders{col},1:vec_size(1)) = row_vals;
    else
        X_final(place_holders{col},(vec_cumsum(col-
1)+1):(vec_cumsum(col))) = row_vals;
    end
end
iterations=iteration;
```

Appendix C: RIMCA algorithm

```
%STEP 1 RIMCA
function [X_final, X_start, X_hat, F_hat,F_hat_dim, L_hat,
U_hat,vec_size]=rimca_new(data,S,threshold)
%data is incomplete (with NaNs)
format long

[I J] = size(data);
X0 = dummyvar(data);
W=1-(isnan(X0)); %obtain weights before filling in initial values for
the missings
vec_size=zeros(J,1);
X = X0;

%Assigning initial values

place_holders = cell(1,J);
for col = 1:J
    vec_size(col)=length(unique(data(isfinite(data(:,col)),col)));
    vec_cumsum = cumsum(vec_size);
    place_holders{col} = find(isnan(data(:,col)));
    if col == 1
      %[row_vals]=initial_vals(vec_size(col),length(place_holders{col}
      ),X0(:,1:vec_size(1))); %SI
```

```
      [row_vals]=initial_vals(vec_size(col),length(place_holders{col})
      ); %MI
      X(place_holders{col},1:vec_size(1)) = row_vals;
    else
      %[row_vals]=initial_vals(vec_size(col),length(place_holders{col}
      ),X0(:,(vec_cumsum(col-1)+1):(vec_cumsum(col)))); %SI

      [row_vals]=initial_vals(vec_size(col),length(place_holders{col})
      ); %MI
X(place_holders{col},(vec_cumsum(col-1)+1):(vec_cumsum(col)))=
row_vals;
    end
end

X_start=X;
[~,K]=size(X);
D_sigma=diag(sum(X));
X_hat = X;

%I = # individuals; J = # of variables, K = # of categories
iteration=0;
iterations=100;
change = 1;
conv_prog=(zeros(sum(sum(isnan(data))),max(vec_size),iterations));
conv_prog_plot=zeros(iterations,
max(vec_size),(sum(sum(isnan(data)))));
%Step 3 RIMCA (while loop)
while (change > threshold && iteration< iterations)
    %Step 2a RIMCA
    row_counter=0;
    Z = I*(X_hat)*(inv(D_sigma));
    M = (1/(I*J))*D_sigma;
    X_hat_reg_shrink = zeros(I,K);
    D=(1/I)*eye(I);

    Z_est=(Z-ones(I,1)*(ones(K,1))')*sqrt(M);
    Z_tilda=D.^(0.5)*Z_est*M.^(0.5);

    [P L Q]=svd(Z_tilda);
    C=D^(-0.5)*P;
    U_hat=M^(-0.5)*Q;
    F_hat=C*L;
    L_hat=L;
    E=(L_hat).^2;

    L_hat_dim=L_hat(1:S,1:S);
    F_hat_dim=F_hat(:,1:S);
    U_hat_dim=U_hat(:,1:S);

    %Step 2b rewritten noise variance REGULARISED RIMCA
    sigma_hat_2=1/(K-J-S)*sum(diag(E((S+1):end,(S+1):end)));
    for i=1:I;
        for k=1:K;
            a=(D_sigma(k,k)/I);
            dist_vec = (sum((F_hat_dim(:,1:S).^2))).^0.5;
shrink=(1./dist_vec(1:S)).*((diag(L_hat_dim(1:S,1:S)))-sigma_hat_2./
(diag(L_hat_dim(1:S,1:S))))';
```

```
outer_sum_shrink=sum(shrink.*(F_hat_dim(i,(1:S)).*U_hat_dim(k,(1:S))))
;
X_hat_reg_shrink(i,k) = a*(outer_sum_shrink+1);
        end
    end
    change = sum(sum(((1-W).*X_hat - (1-W).*X_hat_reg_shrink).^2));
    change_matrix = ((1-W).*X_hat - (1-W).*X_hat_reg_shrink).^2;
    X_hat=(W.*X)+((1-W).*X_hat_reg_shrink);
    D_sigma=diag(sum(X_hat));
    iteration=iteration+1;
    for col = 1:J
        if col == 1;
            fuzzy_vectors = X_hat(place_holders{col},1:vec_size(1));
change_vectors=change_matrix(place_holders{col},1:vec_size(1));
        else
fuzzy_vectors=X_hat(place_holders{col},(vec_cumsum(col-
1)+1):(vec_cumsum(col)));
change_vectors = change_matrix(place_holders{col},(vec_cumsum(col-
1)+1):(vec_cumsum(col)));
        end
    end
end
for rep=1:5
X_final = X_hat;
for col = 1:J
    if col == 1;
        fuzzy_vectors = X_hat(place_holders{col},1:vec_size(1));
    else
        fuzzy_vectors = X_hat(place_holders{col},(vec_cumsum(col-
1)+1):(vec_cumsum(col)));
    end
    %[row_vals] = assign01_to_rows(fuzzy_vectors);%SI
    [row_vals] = assign01_to_rows(fuzzy_vectors,1); %MI
    if col == 1;
        X_final(place_holders{col},1:vec_size(1)) = row_vals;
    else
        X_final(place_holders{col},(vec_cumsum(col-
1)+1):(vec_cumsum(col)))=row_vals;
    end
end
    [data_final] = categorical_recon(X_final, vec_size);
end
iterations=iteration;
```

Appendix D: Simulation Protocol

Data simulation

```
function [output_l,output_h]=simdata(n,p,r_high,r_low,block1,num_cat)

if num_cat == 1
    disp('Can''t have 1 category - set category count as 2')
    num_cat = 2;
end
block2 = p-block1;
bins = ones(1,num_cat);

%low structure correlation
```

```
sigma1_l=r_low*ones(block1,block1)+diag((1-r_low)*ones(block1,1));
sigma2_l=r_low*ones(block2,block2)+diag((1-r_low)*ones(block2,1));
sigma = zeros(block1,block2);
sigma_l=[sigma1_l sigma;sigma' sigma2_l];
output_l=mvnrnd(zeros(p,1),sigma_l,n);
[B IX] = sort(output_l);
for i = 2:num_cat;
    bins(i) = floor(((i-1)*n)/num_cat); % equal count categories
    B(bins(i-1):bins(i),:) = i-1;
end
B((bins(i)+1):end,:)=num_cat; %final category
[~,I] = sort(IX);
for i = 1:p %unsort each column
   output_l(:,i) = B(I(:,i),i);
end

%high structure correlation
sigma1_h=r_high*ones(block1,block1)+diag((1-r_high)*ones(block1,1));
sigma2_h=r_high*ones(block2,block2)+diag((1-r_high)*ones(block2,1));
sigma = zeros(block1,block2);
sigma_h=[sigma1_h sigma;sigma' sigma2_h];
output_h=mvnrnd(zeros(p,1),sigma_h,n);
[B IX] = sort(output_h);
for i = 2:num_cat;
    bins(i) = floor(((i-1)*n)/num_cat); % equal count categories
    B(bins(i-1):bins(i),:) = i-1;
end
B((bins(i)+1):end,:)=num_cat; %final category
[~,I] = sort(IX);
for i = 1:p %unsort each column
   output_h(:,i) = B(I(:,i),i);
end
```

MAR and MCAR mechanisms

```
function [MCAR_LR, MCAR_LNR, MCAR_HR, MCAR_HNR, MAR_LR, MAR_LNR,
MAR_HR, MAR_HNR] = josseholes(data)
[n p] = size(data);

mcar_1 = 0.1;
mcar_2 = 0.3;
mar_1 = 0.08;
mar_2 = 0.16;

%MCAR
%Case 1 LR 10%
MCAR_LR = data;
while sum(sum(isnan(MCAR_LR)))/(n*p)< mcar_1
    MCAR_LR(randi(100,1),randi(10,1)) = NaN;
end
%Case 2 LNR 10%
MCAR_LNR = data;
MCAR_LNR(1:20,1:3) = NaN(20,3);
MCAR_LNR(81:100,9:10) = NaN(20,2);
%Case 3 HR 30%
MCAR_HR = data;
while sum(sum(isnan(MCAR_HR)))/(n*p)< mcar_2
    MCAR_HR(randi(100,1),randi(10,1)) = NaN;
end
%Case 4 HNR 30%
MCAR_HNR = data;
```

```
MCAR_HNR(1:60,1:3) = NaN(60,3);
MCAR_HNR(41:100,9:10) = NaN(60,2);
%MAR
%Case 5 LR 8%
MAR_LR = data;
while sum(sum(isnan(MAR_LR)))/(n*p) < mar_1
    IX = find(MAR_LR(:,1)==1,randi(33,1));
    MAR_LR(IX(end),2:6) = NaN(1,5);
    IX = find(MAR_LR(:,7)==3,randi(34,1));
    MAR_LR(IX(end),8:10) = NaN(1,3);
end
%Case 6 LNR 8%
MAR_LNR = data;
MAR_LNR(find(MAR_LNR(:,1)==1,10),2:6)=NaN(10,5);
MAR_LNR(find(MAR_LNR(:,7)==3,10,'last'),8:10)=NaN(10,3);
%Case 7 HR 16%
MAR_HR = data;
while sum(sum(isnan(MAR_HR)))/(n*p) < mar_2
    IX = find(MAR_HR(:,1)==1,randi(33,1));
    MAR_HR(IX(end),2:6) = NaN(1,5);
    IX = find(MAR_HR(:,7)==3,randi(34,1));
    MAR_HR(IX(end),8:10) = NaN(1,3);
end
%Case 8 HNR 16%
MAR_HNR = data;
MAR_HNR(find(MAR_HNR(:,1)==1,20),2:6)=NaN(20,5);
MAR_HNR(find(MAR_HNR(:,7)==3,20,'last'),8:10)=NaN(20,3);
```

Appendix E: Code for the selection of 10 random dimensions

```
clear
clc
n=19; %max # dimensions
m=10; %# dimensions
r=randperm(n);
dim_vec=sort(r(1:m))';
```

Appendix F: Code for CI's of singly imputed datasets

Simulated data

```
clear
clc
data_original=xlsread('<path>\filename.xlsx');
data_imputed=xlsread('<path>\filename.xlsx');
[m variable] = size(data_original);
v=m-1;
tvals = tinv(0.975,v);

Q_avg_org=mean(data_original);
Q_variance_org=var(data_original)/m;
Q_avg_imp=mean(data_imputed);
Q_variance_imp=var(data_imputed)/m;

CIlow_org =(Q_avg_org)-(tvals.*((Q_variance_org).^(0.5)));
CIhigh_org =(Q_mean_org)+(tvals.*((Q_variance_org).^(0.5)));
CIlow_imp=(Q_avg_imp)-(tvals.*((Q_variance_imp).^(0.5)));
CIhigh_imp=(Q_avg_imp)+(tvals.*((Q_variance_imp).^(0.5)));
```

```
results_org=[Q_avg_org; Q_variance_org; CIlow_org;CIhigh_org];
results_imp=[Q_avg_imp; Q_variance_imp; CIlow_imp; CIhigh_imp];
```

real data

```
dat_original=xlsread('<path>\filename.xlsx');
data_imputed = xlsread('<path>\filename.xlsx');
[m variable] = size(Q_hat_original);
N=sum(isfinite(Q_hat_original));
v=N-1;
tvals = tinv(0.975,v);

Q_avg_org=nanmean(data_original);
Q_variance_org=nanvar(data_original)/m;

CIlow_org =(Q_avg_org)-(tvals.*((Q_variance_org).^(0.5)));
CIhigh_org =(Q_avg_org)+(tvals.*((Q_variance_org).^(0.5)));

results_org=[Q_mean_org; Q_var_org; CIlow_org;CIhigh_org];
```

Appendix G: Code for Rubin's Rules

```
data_orginal=xlsread('<path>\filename.xlsx');
Q_est=xlsread('<path>\filename.xlsx');
U_est=xlsread('<path>\filename.xlsx');

[n,~]=size(data_original);
[m var]=size(Q_est);
Q_bar=(1/m)*sum(Q_est,1);
U_bar=(1/m)*sum(U_est,1);
B=(1/(m-1))*(sum((Q_est-(ones(m,1)*Q_bar)).^2));

T=U_bar+(1+(1/m)).*B;
r=(1+(1/m)).*(B./U_bar);
v=(m-1).*(1+(1./r)).^2;
rmi=(r+2./(v+3))./(r+1);

eff=(1+rmi/m).^(-0.5);
tvals=tinv(0.975,v);
CIlow=Q_bar-((T.^0.5).* tvals);
CIhigh=Q_bar+((T.^0.5).* tvals);

results=[Q_bar; U_bar; B; T; v; rmi; r; eff; tvals; CIlow; CIhigh];
```

Appendix H: Code for Apparent Error Rate

```
clear
clc
data_original=xlsread('<path>\filename.xlsx');
data_imputed=xlsread('<path>\filename.xlsx');
x=dummyvar(data_original);
x_hat=dummyvar(data_imputed);
[I J] =size(data_original);
diff=x-x_hat;
err=(sum(sum(abs(diff)))/2)/(I*J);
rate=err*100;
```

Appendix I: Description of the variables of the user satisfaction survey: *Canal des Deux Mers*

Var	Name	Variable description	Categories
1	Sites worth visiting	What do you think about information given about the sights worth visiting?	Satisfactory, unsatisfactory, no opinion
2	Leisure activity	Rate the information given on leisure activity.	Satisfactory, unsatisfactory, no opinion
3	Historical canal sites	What is your opinion regarding tourist information on historical canal sites (docks, bridges, etc.)?	Satisfactory, unsatisfactory, no opinion
4	Manoeuvres	Where you sufficiently aware of manoeuvres at docks at the start of your trip?	Yes, no
5	Authorized mooring	Where you sufficiently aware of authorized mooring at the start of your trip?	Yes, no
6	Safety regulations	Where you sufficiently aware of safety regulations at the start of your trip?	Yes, no
7	Services	What is your opinion on signs encountered along the way providing information on services?	Satisfactory, unsatisfactory
8	Number of laps	What do you think about the number of laps on your trip?	Sufficient, insufficient
9	Cost of water	The general cost of water is…	Inexpensive, average, expensive
10	Cost of electricity	The general cost of electricity is…	Inexpensive, average, expensive
11	Visibility of electrical outlets	What is your opinion of visibility of electrical outlets?	Sufficient, insufficient
12	Number of electrical outlets	What do you think about the number of electrical outlets on your trip?	Sufficient, insufficient
13	Cleanliness	How would you describe the canal's degree of cleanliness?	Clean, average, dirty
14	Unpleasant odours	Were there unpleasant odours on the canal?	None, occasional, frequent

(Chavent, Kuentz and Saracco 2010:97)

Appendix J: Number of iterations before the algorithm in question converges over all dimensions

Algorithm	Number of iterations
RIMCA	9
IMCA	9,10,**11**,12

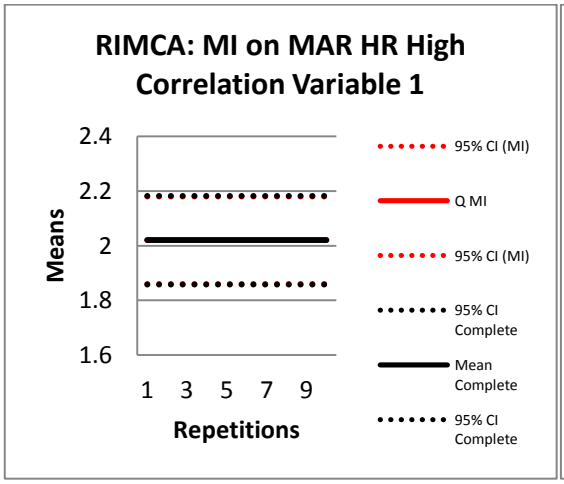
The Bold font indicates the majority of iterations for the case where more than one iteration value was obtained.

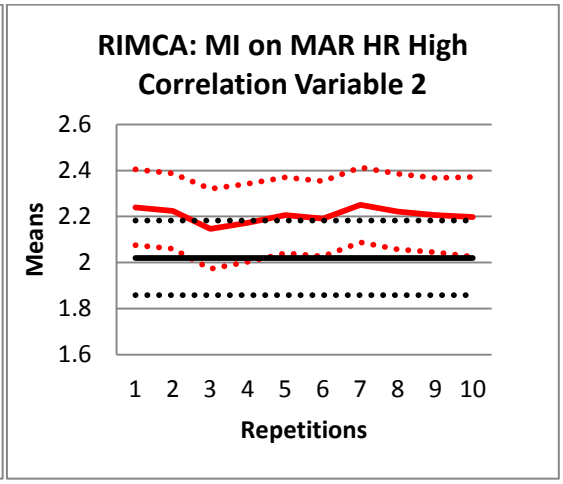
Number of iterations before algorithm converge over all dimensions								
Correlation Structure	High percentage of missing values				Low percentage of missing values			
	Random		Non-random		Random		Non-random	
	MAR	MCAR	MAR	MCAR	MAR	MCAR	MAR	MCAR
High	7	**9**, 10	7	**16**, 17	5	5	5	6
Low	6, **7**	**8**, 9	6, **7**	**16**, 17	5	5	5	6

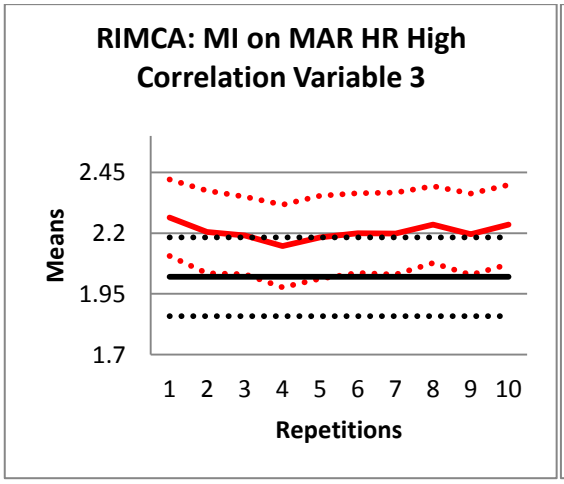
The Bold font indicates the majority of iterations for the case where more than one iteration value was obtained.

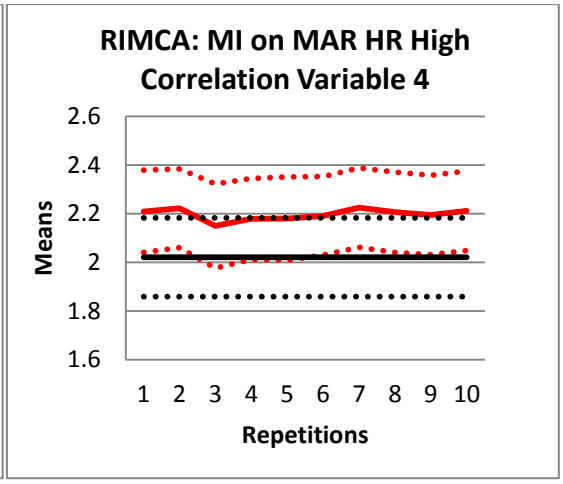
Appendix K: Stability graphs over ten repetitions

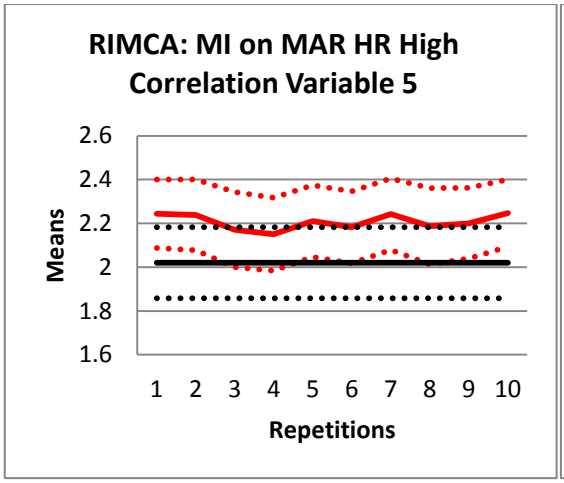
K.1 MAR random pattern with 16% missing values and high correlation structure

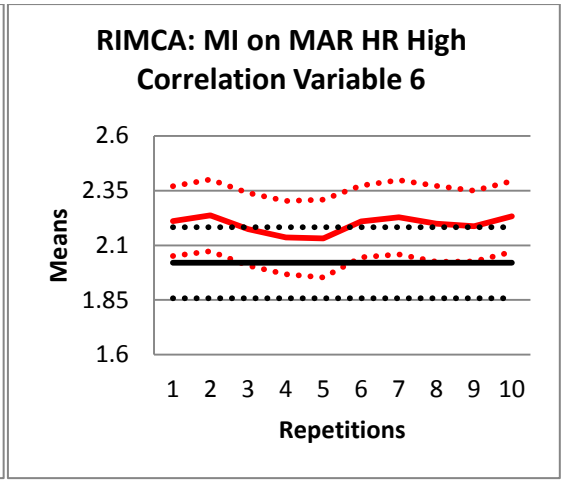


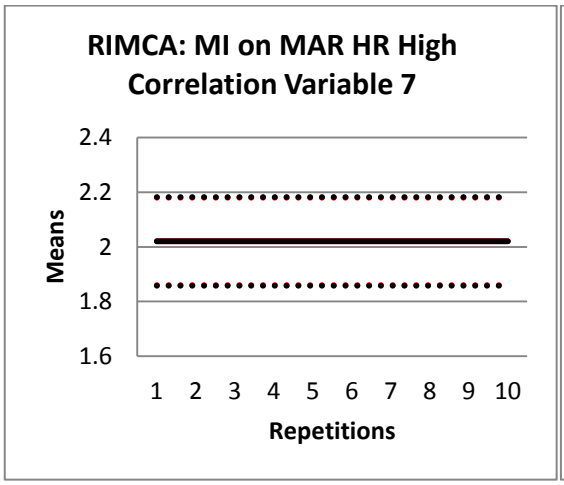


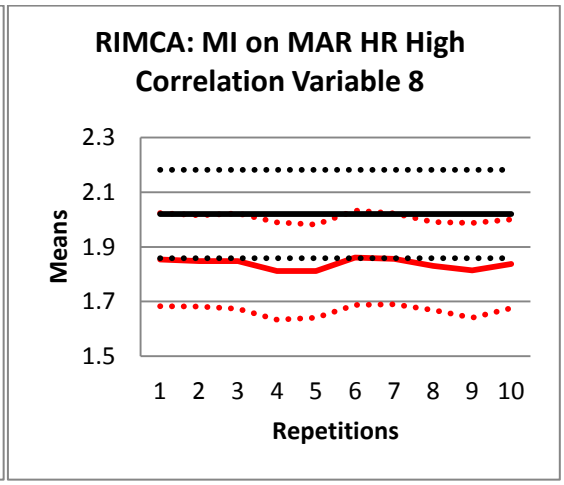


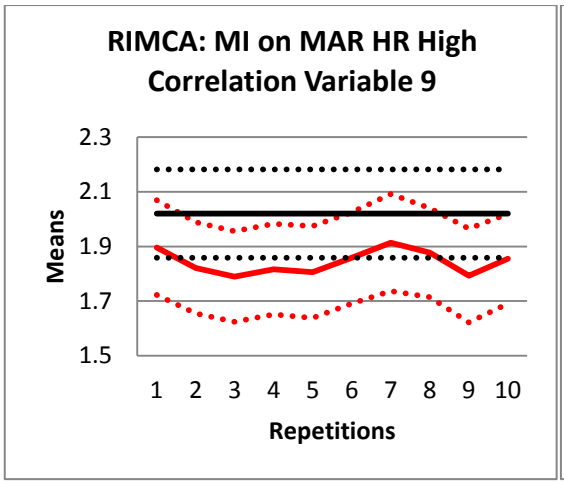


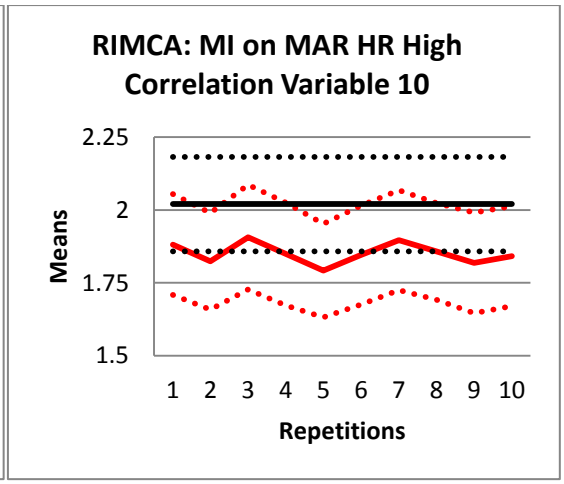

RIMCA: MI on MAR HR High Correlation Variable 5
Means
2.6
2.4
2.2
2
1.8
1.6
1 2 3 4 5 6 7 8 9 10
Repetitions


RIMCA: MI on MAR HR High Correlation Variable 6
Means
2.6
2.35
2.1
1.85
1.6
1 2 3 4 5 6 7 8 9 10
Repetitions

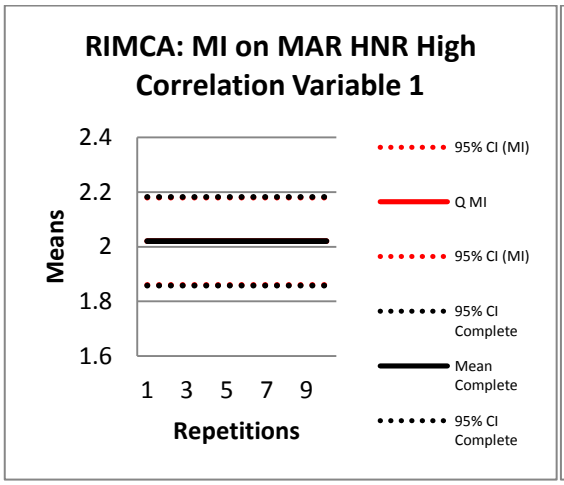

RIMCA: MI on MAR HR High Correlation Variable 7
Means
2.4
2.2
2
1.8
1.6
1 2 3 4 5 6 7 8 9 10
Repetitions

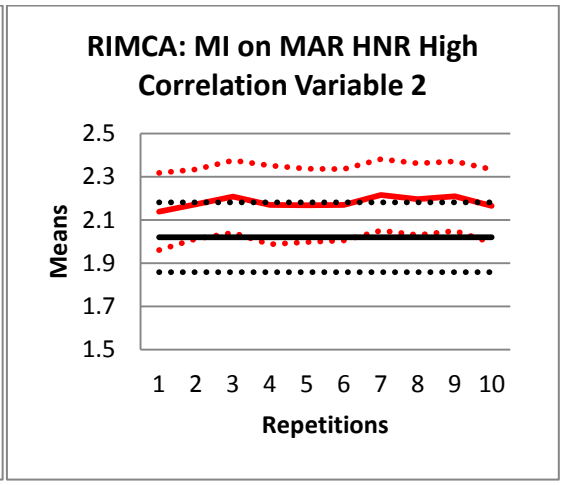

RIMCA: MI on MAR HR High Correlation Variable 8
Means
2.3
2.1
1.9
1.7
1.5
1 2 3 4 5 6 7 8 9 10
Repetitions

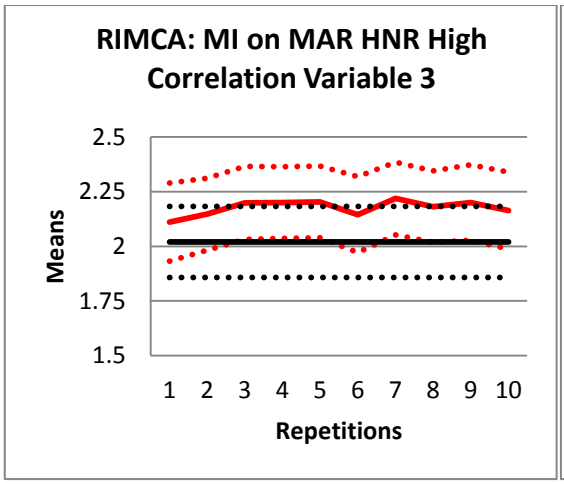

RIMCA: MI on MAR HR High Correlation Variable 9
Means
2.3
2.1
1.9
1.7
1.5
1 2 3 4 5 6 7 8 9 10
Repetitions

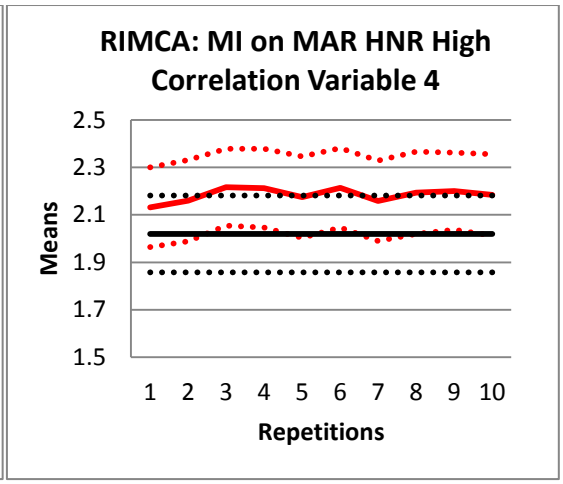

RIMCA: MI on MAR HR High Correlation Variable 10
Means
2.25
2
1.75
1.5
1 2 3 4 5 6 7 8 9 10
Repetitions

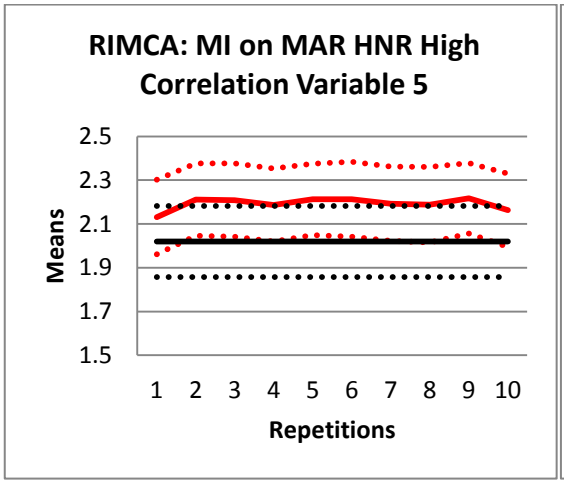
K.2 MAR non-random pattern with 16% missing values and high correlation structure

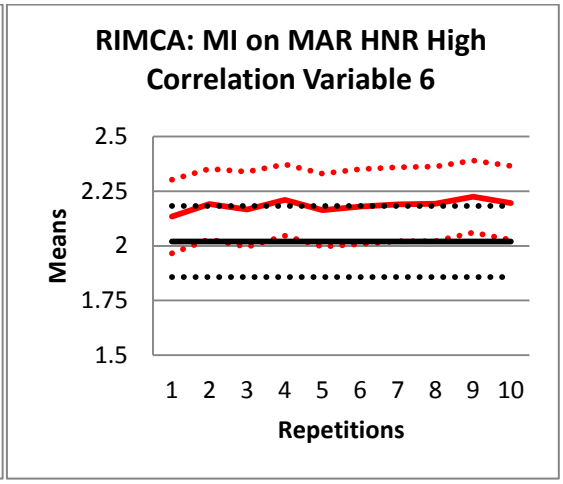


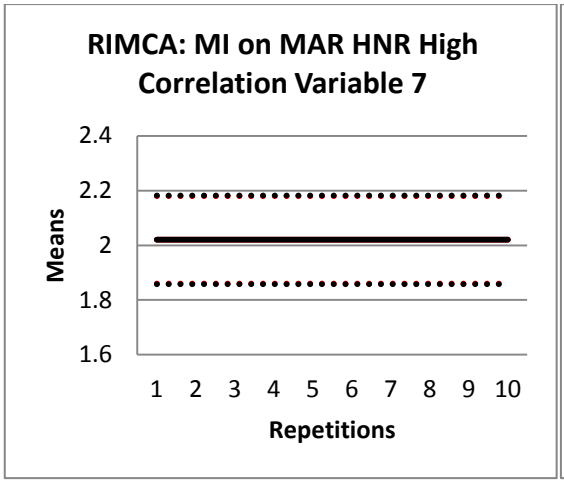


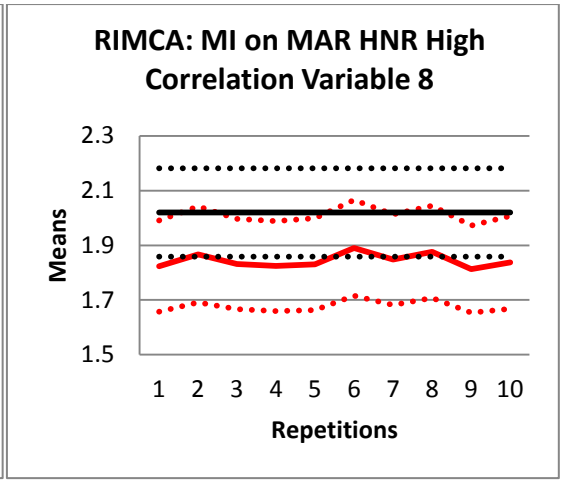


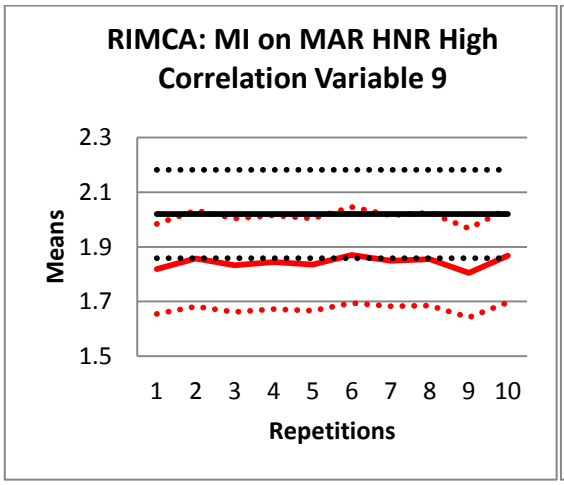


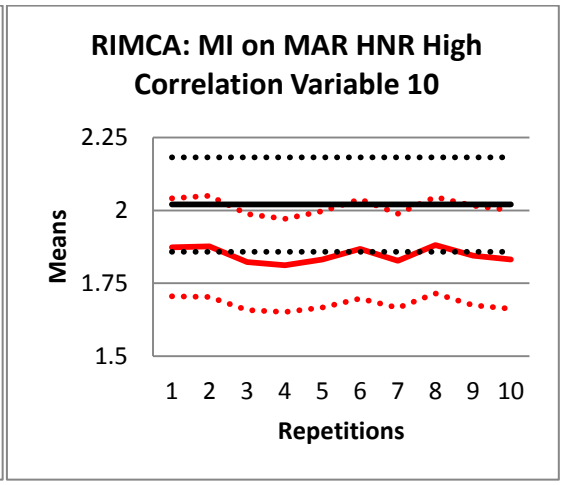




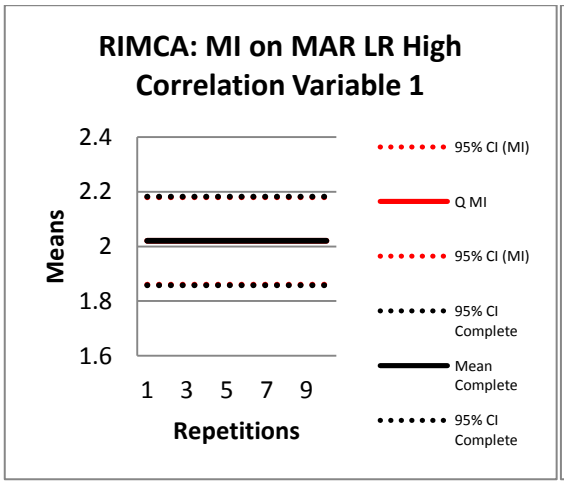


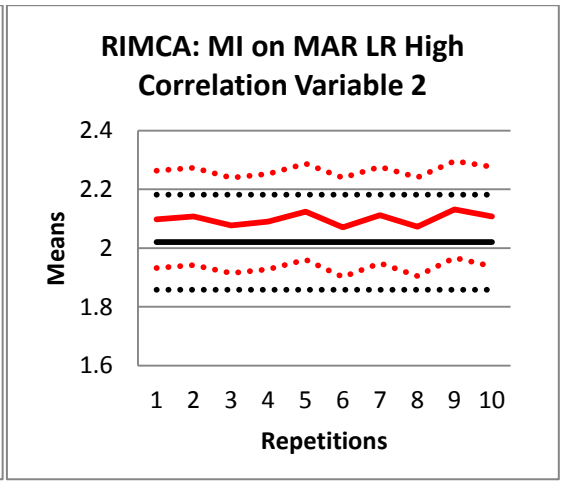


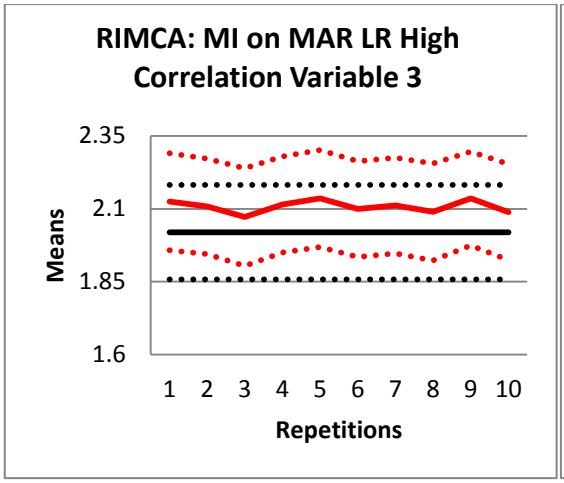


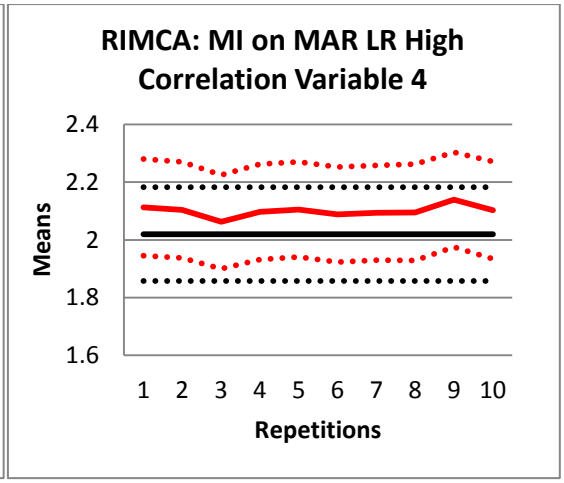


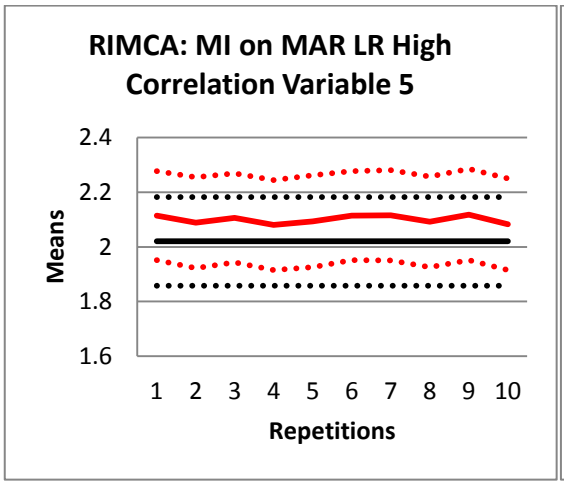
K.3 MAR random pattern with 8% missing values and high correlation structure

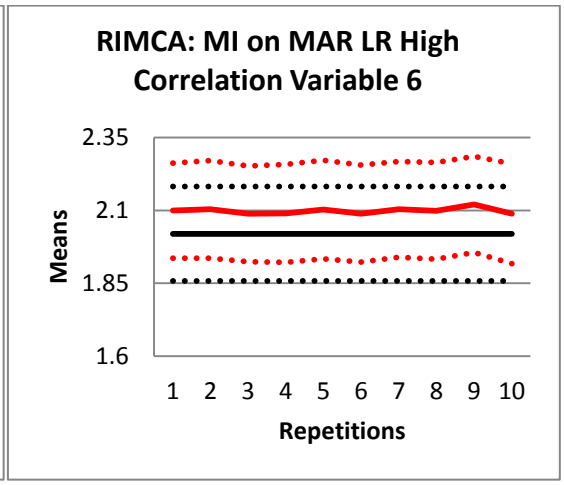


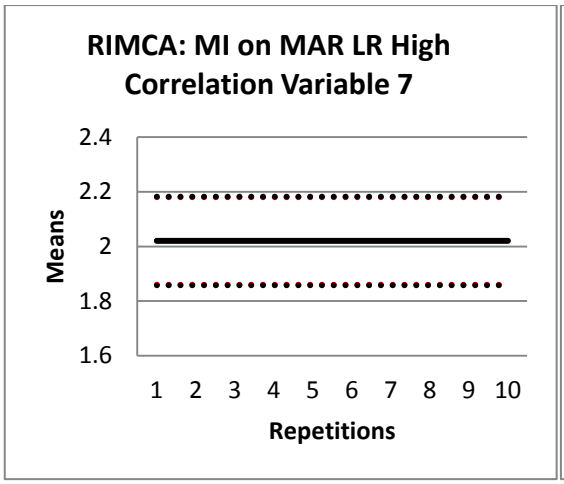


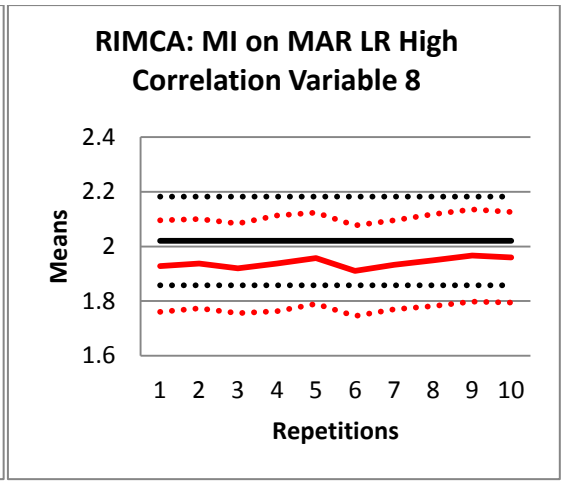

RIMCA: MI on MAR LR High Correlation Variable 3
Means
2.35
2.1
1.85
1.6
1 2 3 4 5 6 7 8 9 10
Repetitions

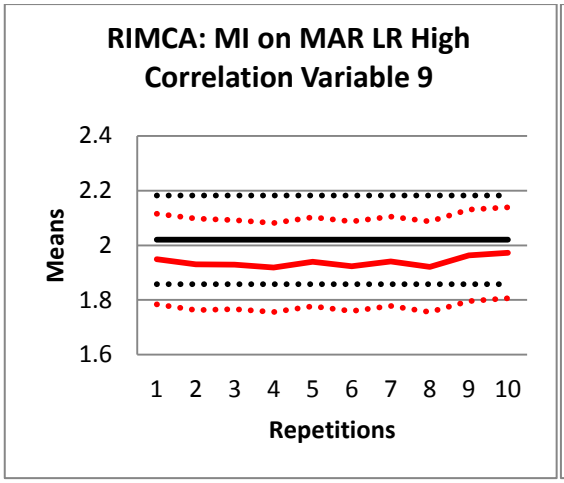

RIMCA: MI on MAR LR High Correlation Variable 4
Means
2.4
2.2
2
1.8
1.6
1 2 3 4 5 6 7 8 9 10
Repetitions

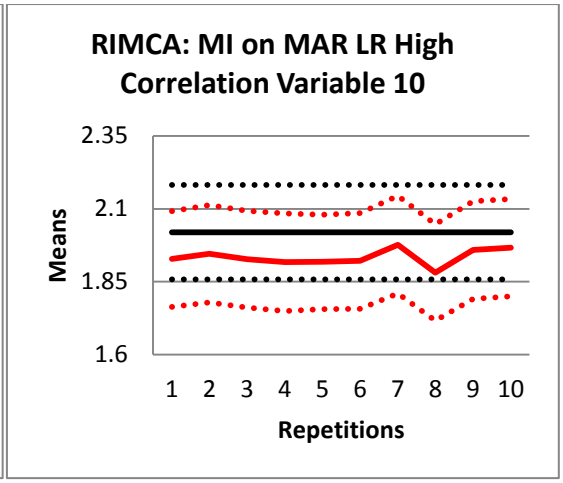

RIMCA: MI on MAR LR High Correlation Variable 5
Means
2.4
2.2
2
1.8
1.6
1 2 3 4 5 6 7 8 9 10
Repetitions


RIMCA: MI on MAR LR High Correlation Variable 6
Means
2.35
2.1
1.85
1.6
1 2 3 4 5 6 7 8 9 10
Repetitions

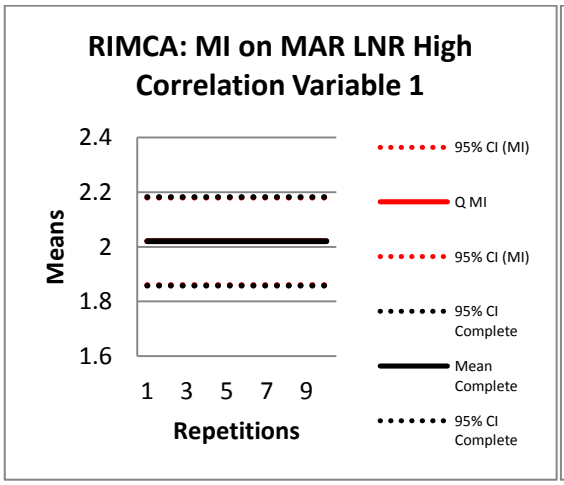

RIMCA: MI on MAR LR High Correlation Variable 7
Means
2.4
2.2
2
1.8
1.6
1 2 3 4 5 6 7 8 9 10
Repetitions

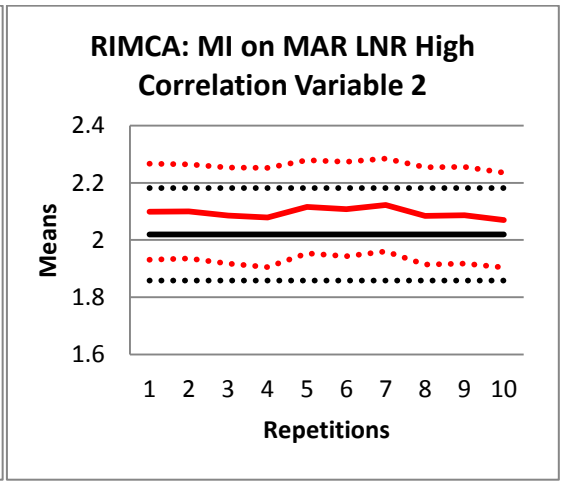

RIMCA: MI on MAR LR High Correlation Variable 8
Means
2.4
2.2
2
1.8
1.6
1 2 3 4 5 6 7 8 9 10
Repetitions

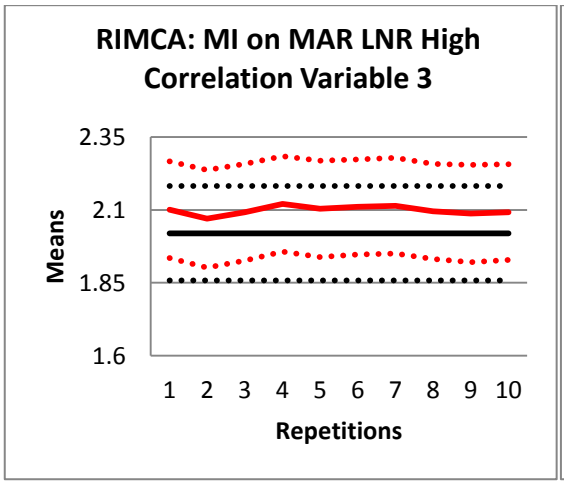


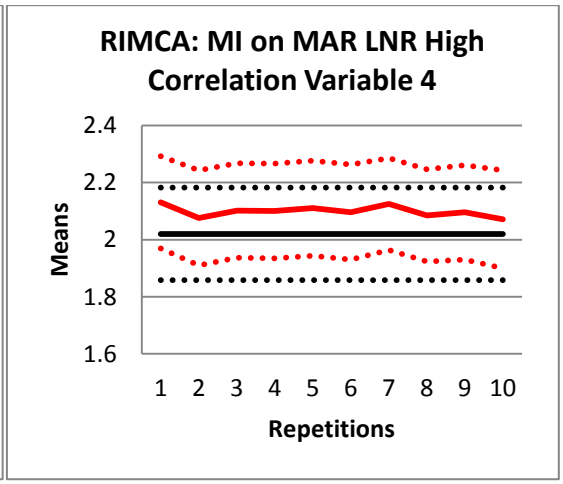


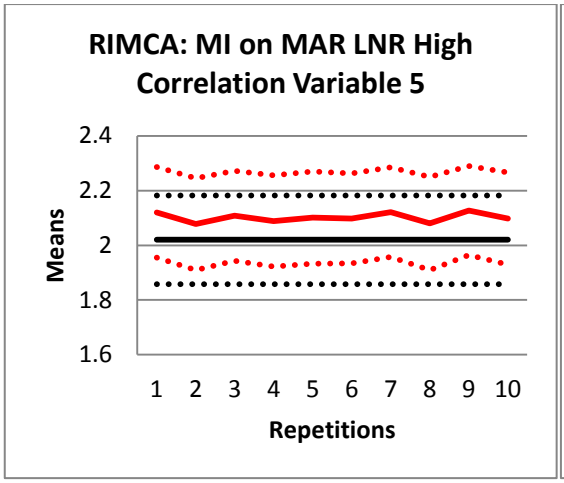
K.4 MAR non-random pattern with 8% missing values and high correlation structure

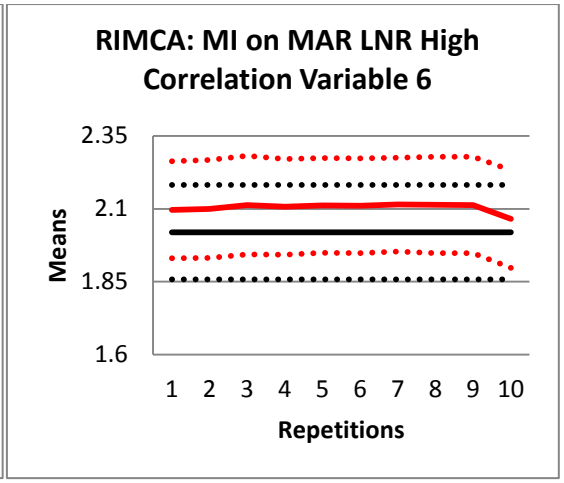


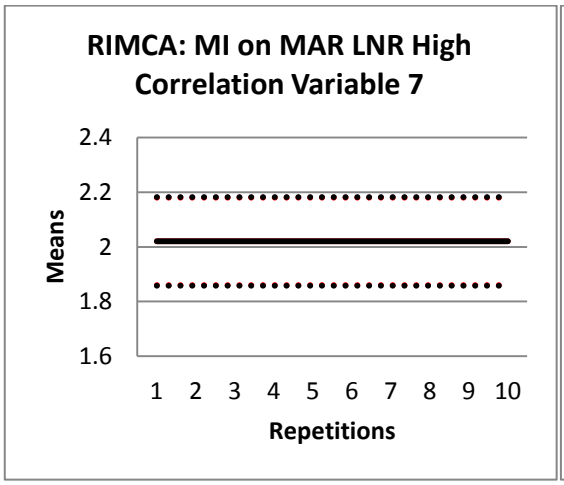


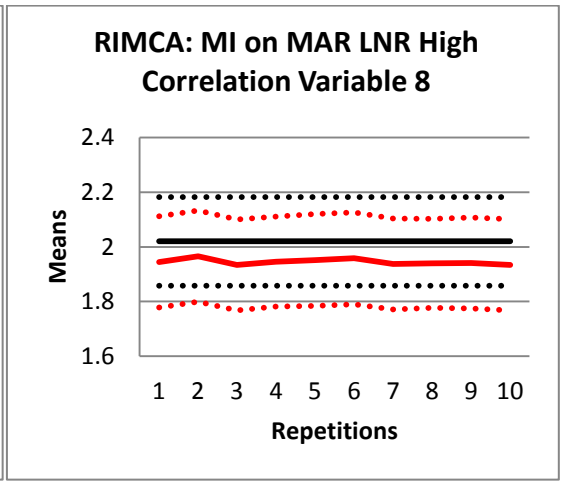


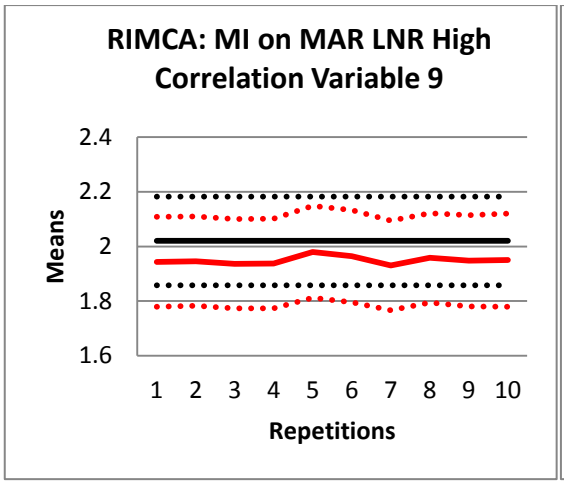


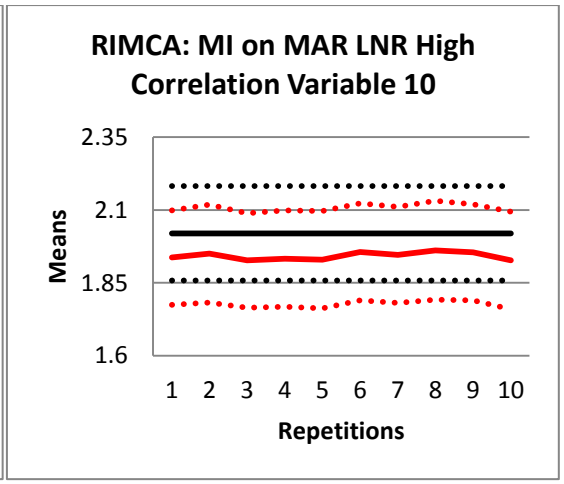

RIMCA: MI on MAR LNR High Correlation Variable 5
Means
2.4
2.2
2
1.8
1.6
1 2 3 4 5 6 7 8 9 10
Repetitions


RIMCA: MI on MAR LNR High Correlation Variable 6
Means
2.35
2.1
1.85
1.6
1 2 3 4 5 6 7 8 9 10
Repetitions

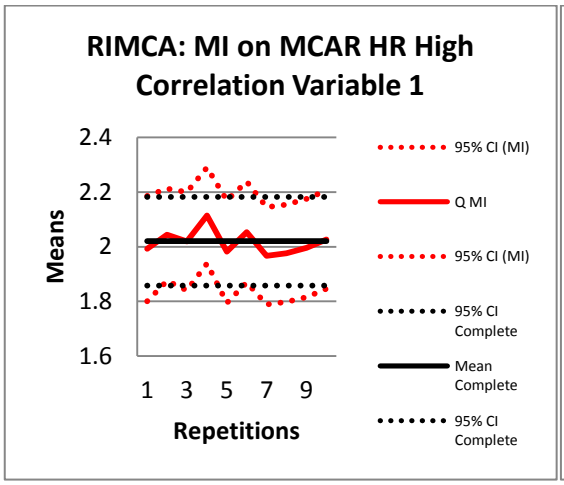

RIMCA: MI on MAR LNR High Correlation Variable 7
Means
2.4
2.2
2
1.8
1.6
1 2 3 4 5 6 7 8 9 10
Repetitions

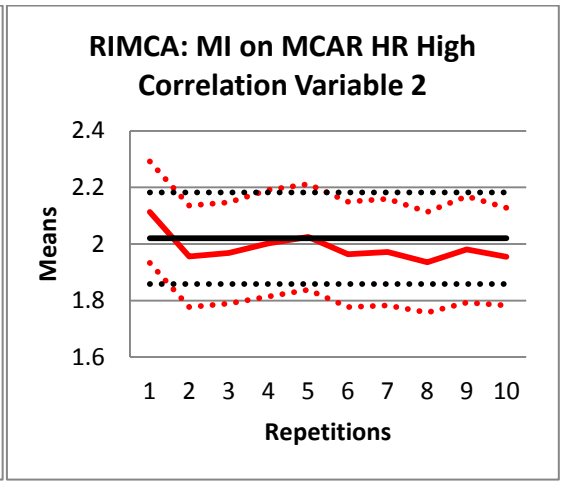

RIMCA: MI on MAR LNR High Correlation Variable 8
Means
2.4
2.2
2
1.8
1.6
1 2 3 4 5 6 7 8 9 10
Repetitions

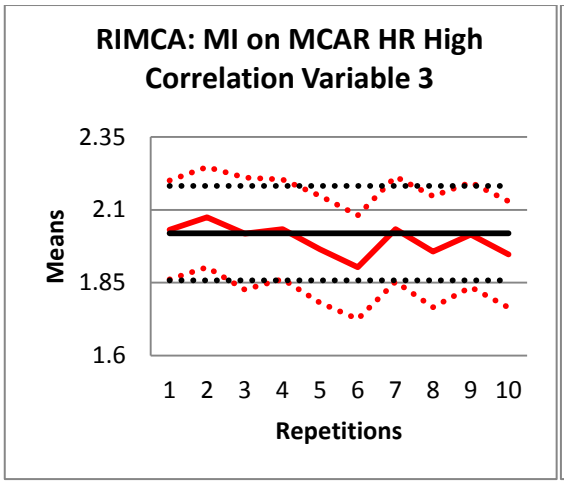

RIMCA: MI on MAR LNR High Correlation Variable 9
Means
2.4
2.2
2
1.8
1.6
1 2 3 4 5 6 7 8 9 10
Repetitions

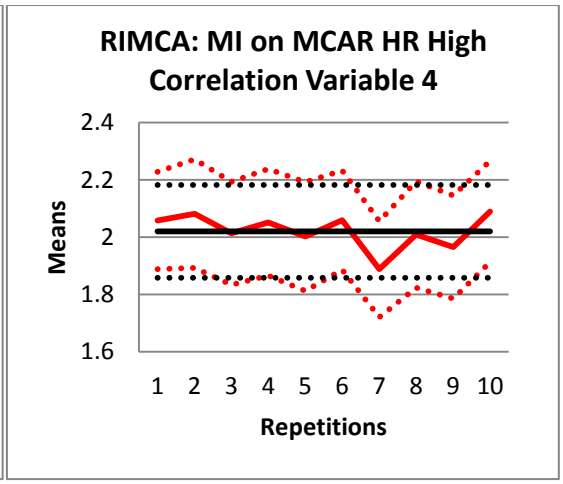

RIMCA: MI on MAR LNR High Correlation Variable 10
Means
2.35
2.1
1.85
1.6
1 2 3 4 5 6 7 8 9 10
Repetitions

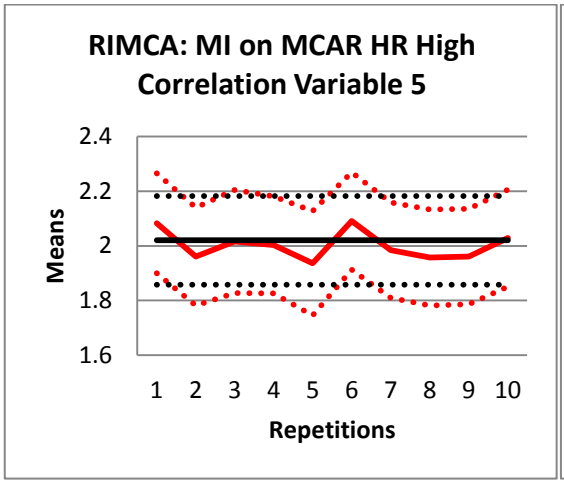
K.5 MCAR random pattern with 30% missing values and high correlation structure

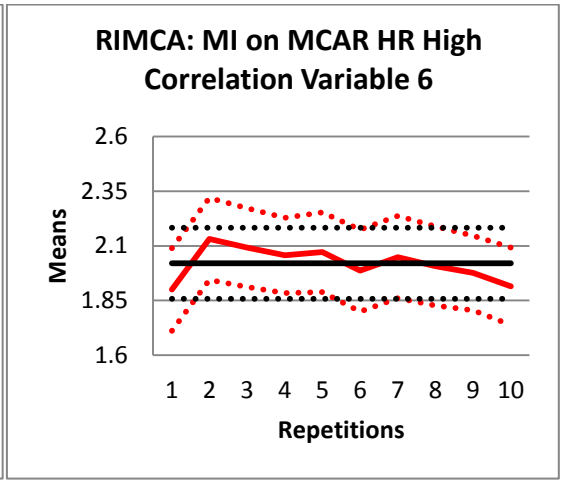


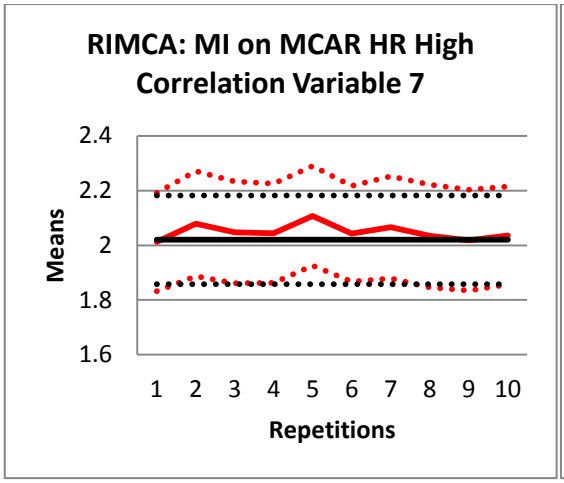


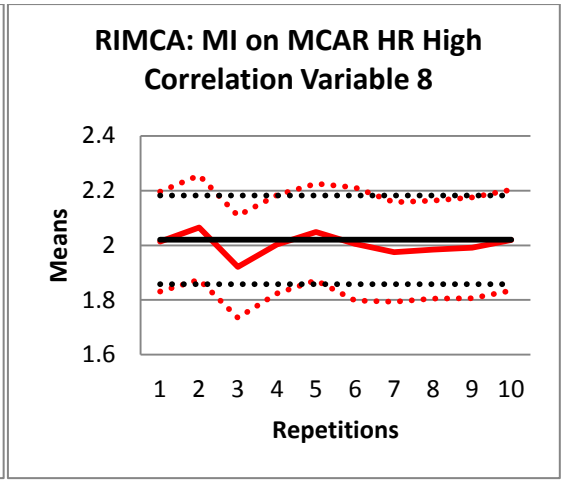


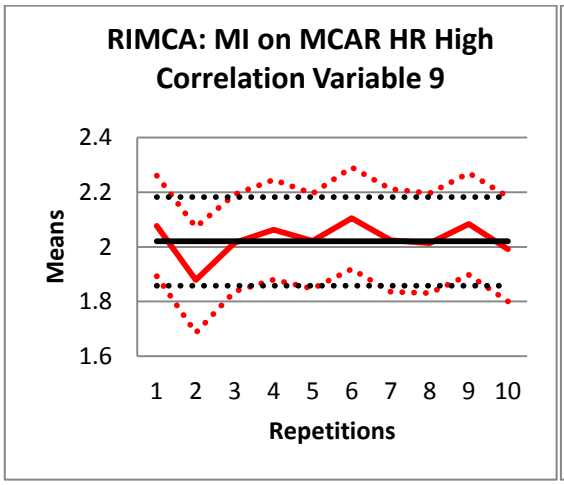


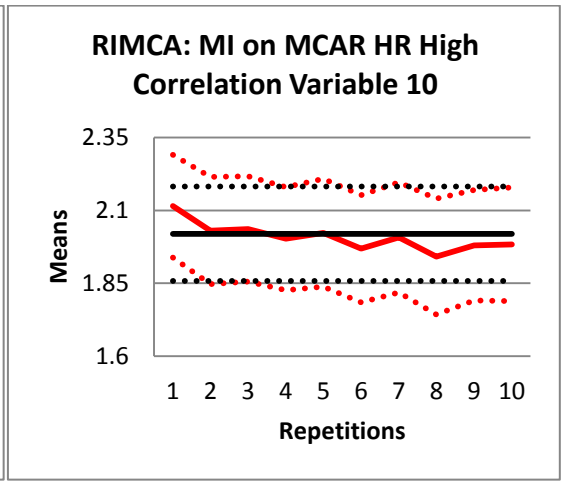




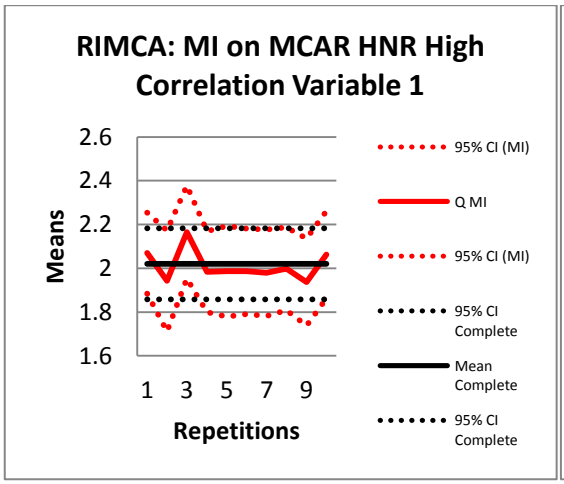


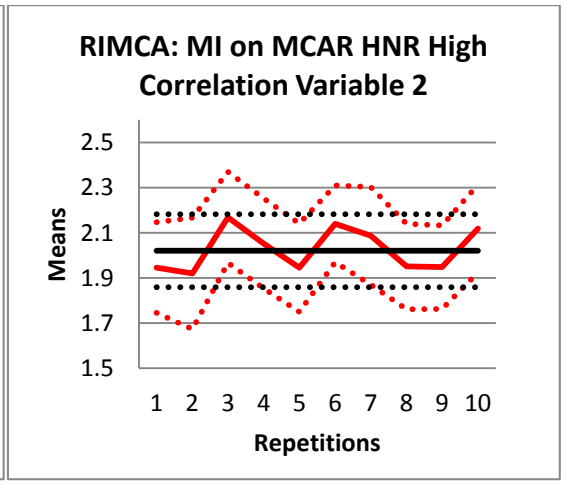


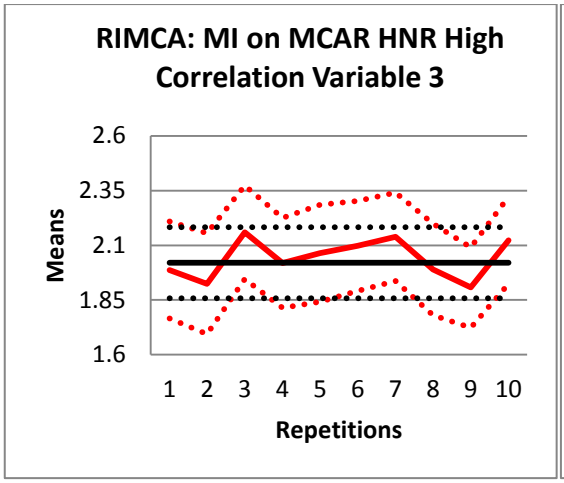


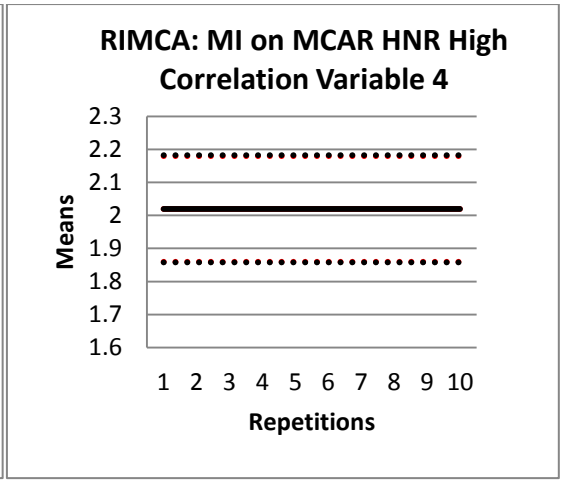


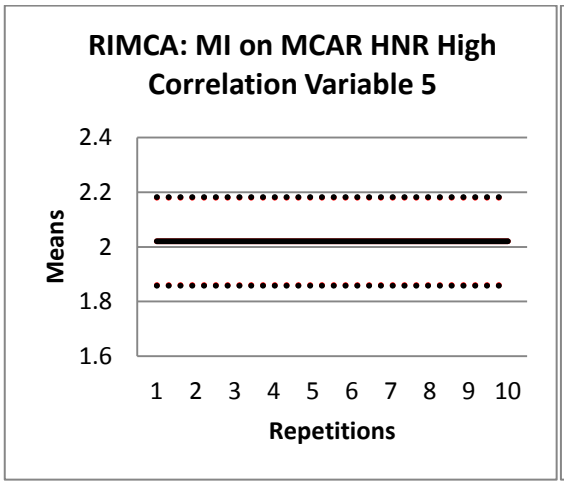
K.6 MCAR non-random pattern with 30% missing values and high correlation structure

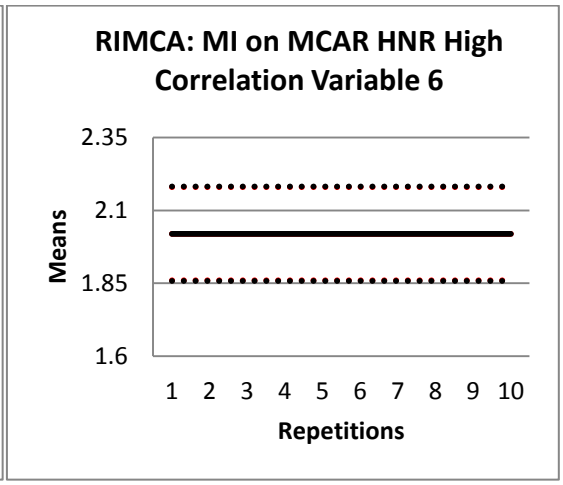


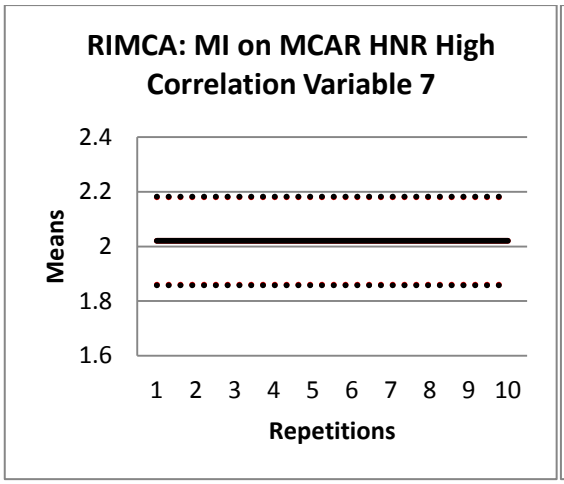


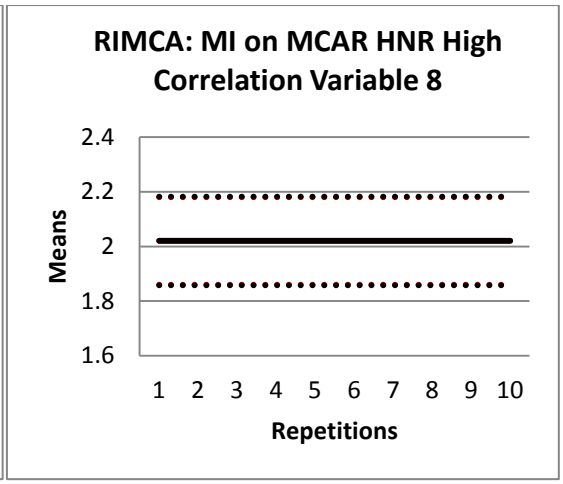

RIMCA: MI on MCAR HNR High Correlation Variable 3
Means
2.6
2.35
2.1
1.85
1.6
1 2 3 4 5 6 7 8 9 10
Repetitions

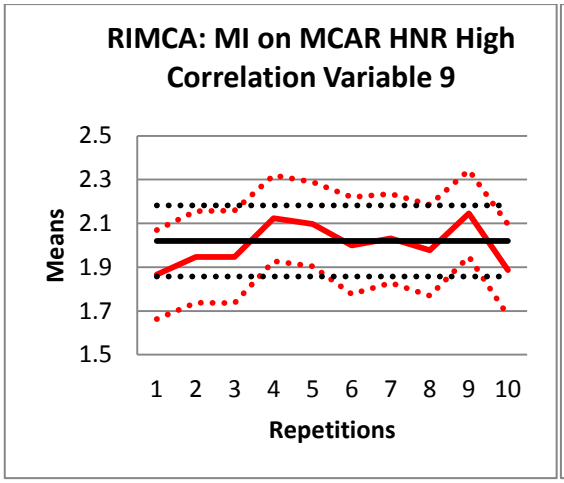

RIMCA: MI on MCAR HNR High Correlation Variable 4
Means
2.3
2.2
2.1
2
1.9
1.8
1.7
1.6
1 2 3 4 5 6 7 8 9 10
Repetitions

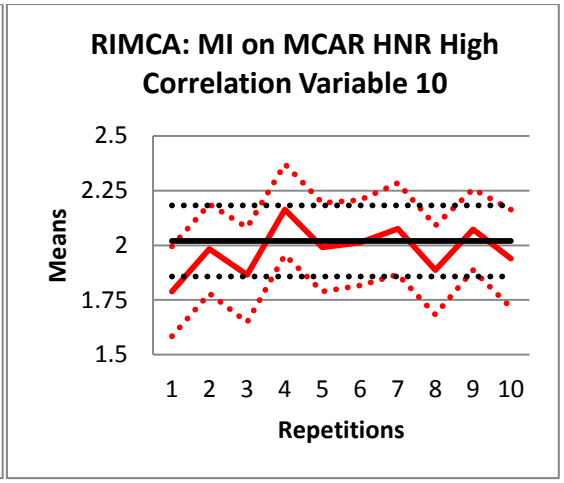

RIMCA: MI on MCAR HNR High Correlation Variable 5
Means
2.4
2.2
2
1.8
1.6
1 2 3 4 5 6 7 8 9 10
Repetitions


RIMCA: MI on MCAR HNR High Correlation Variable 6
Means
2.35
2.1
1.85
1.6
1 2 3 4 5 6 7 8 9 10
Repetitions

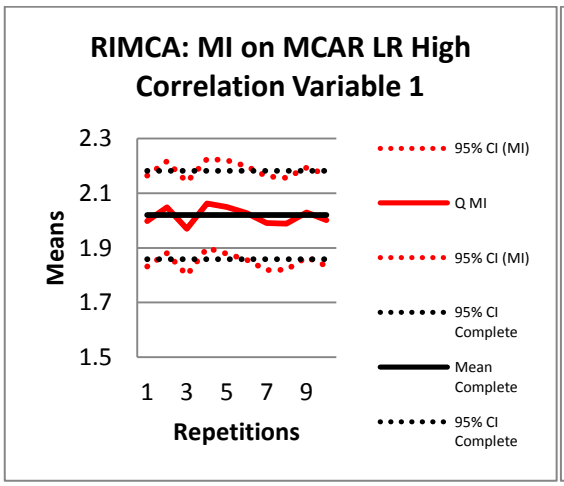

RIMCA: MI on MCAR HNR High Correlation Variable 7
Means
2.4
2.2
2
1.8
1.6
1 2 3 4 5 6 7 8 9 10
Repetitions

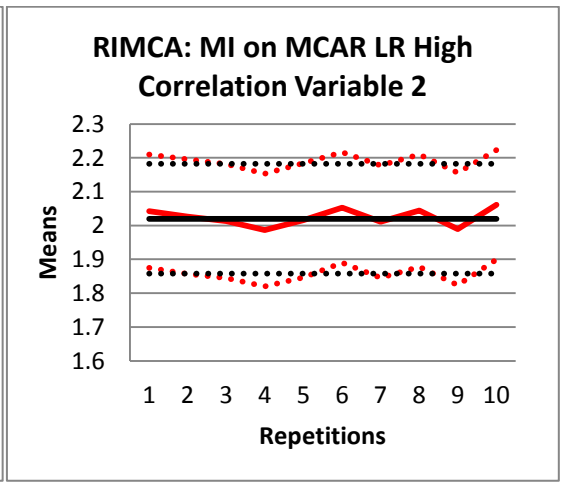

RIMCA: MI on MCAR HNR High Correlation Variable 8
Means
2.4
2.2
2
1.8
1.6
1 2 3 4 5 6 7 8 9 10
Repetitions

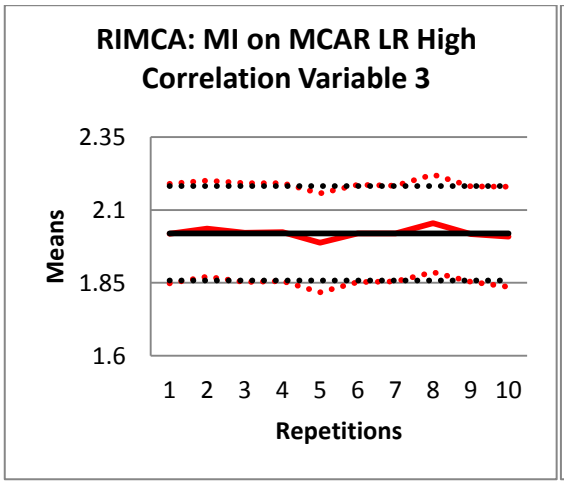


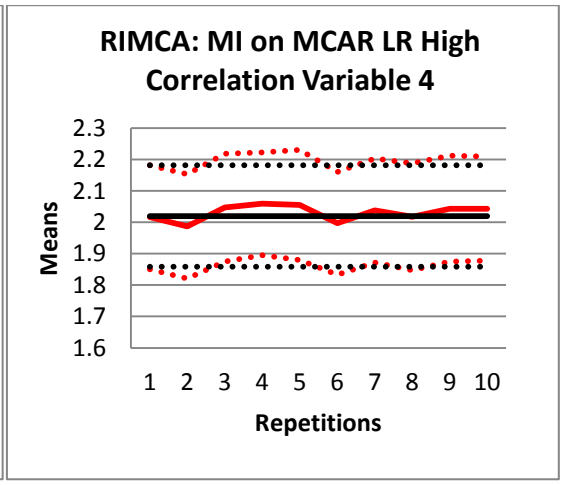


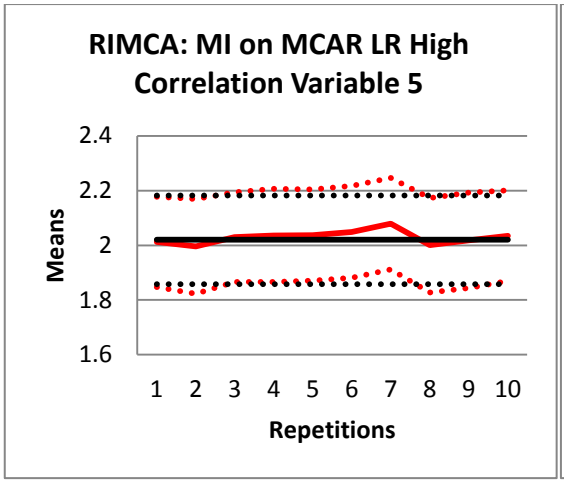
K.7 MCAR random pattern with 10% missing values and high correlation structure

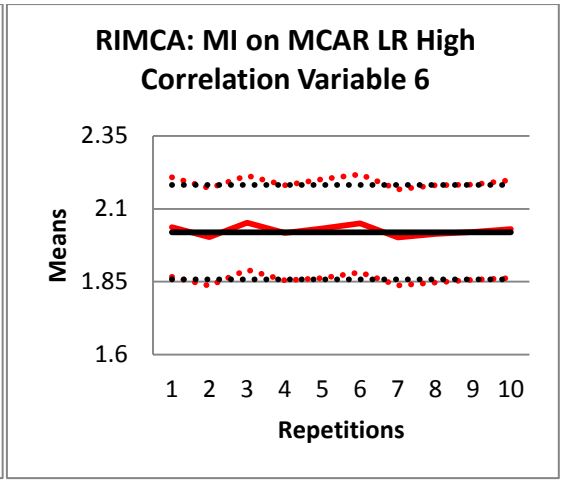


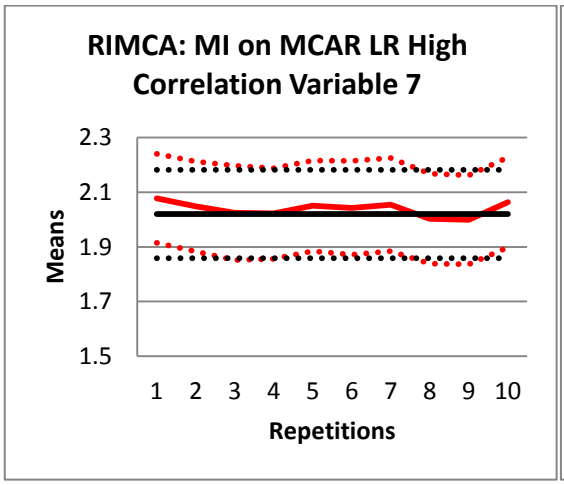


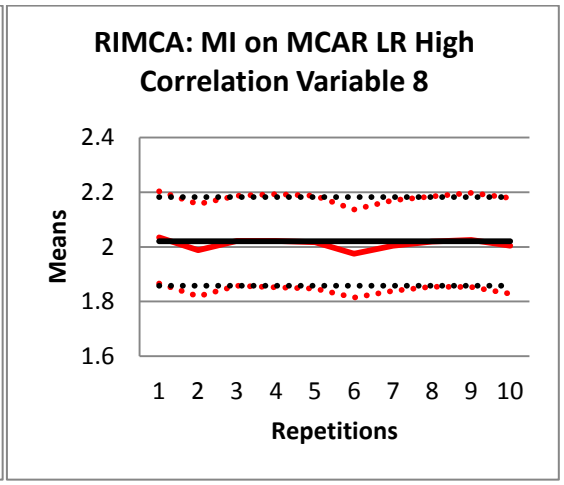


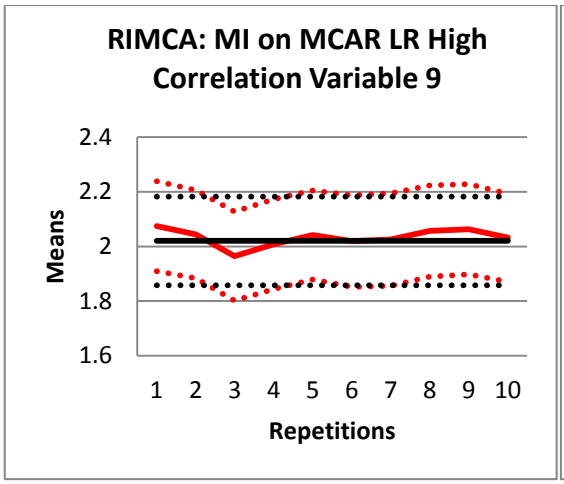


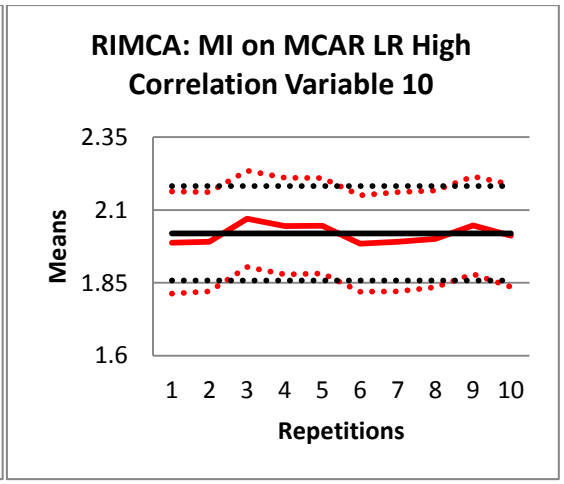

RIMCA: MI on MCAR LR High Correlation Variable 5
Means
2.4
2.2
2
1.8
1.6
1 2 3 4 5 6 7 8 9 10
Repetitions


RIMCA: MI on MCAR LR High Correlation Variable 6
Means
2.35
2.1
1.85
1.6
1 2 3 4 5 6 7 8 9 10
Repetitions

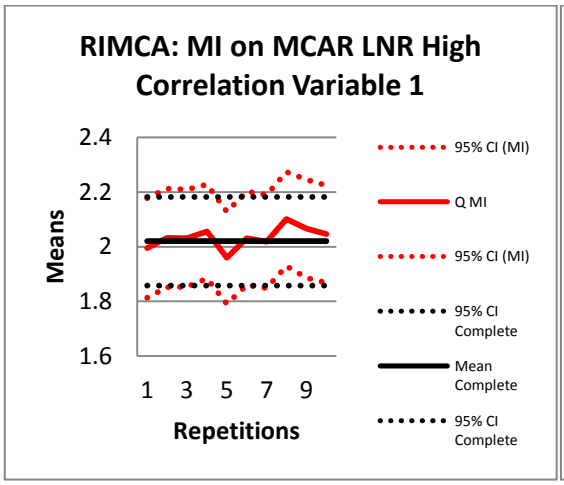

RIMCA: MI on MCAR LR High Correlation Variable 7
Means
2.3
2.1
1.9
1.7
1.5
1 2 3 4 5 6 7 8 9 10
Repetitions

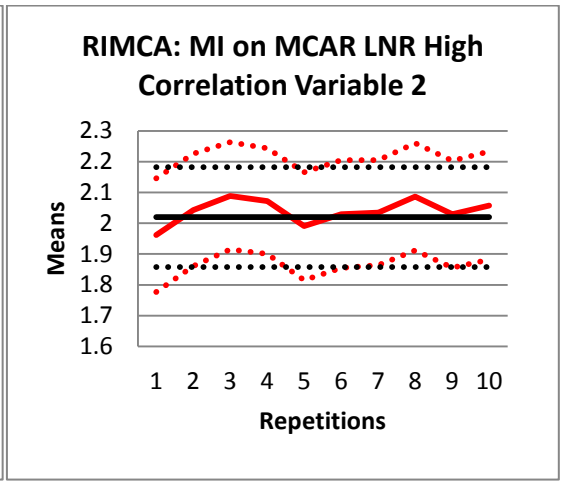

RIMCA: MI on MCAR LR High Correlation Variable 8
Means
2.4
2.2
2
1.8
1.6
1 2 3 4 5 6 7 8 9 10
Repetitions

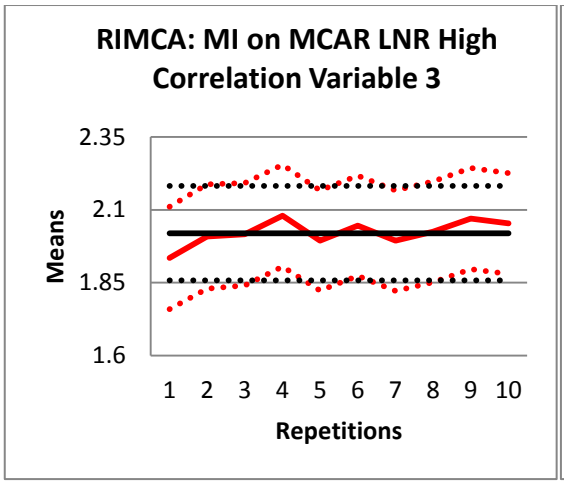

RIMCA: MI on MCAR LR High Correlation Variable 9
Means
2.4
2.2
2
1.8
1.6
1 2 3 4 5 6 7 8 9 10
Repetitions

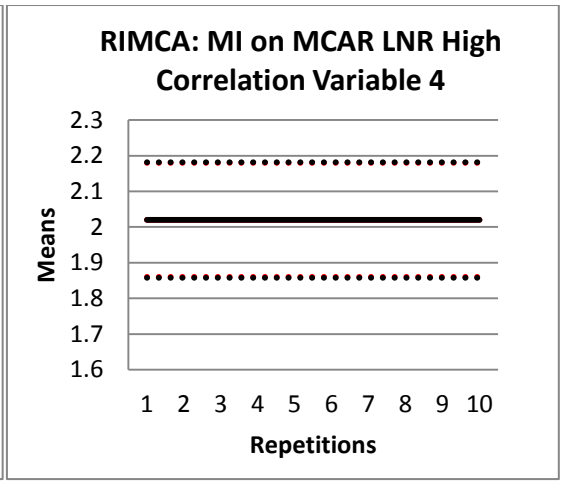

RIMCA: MI on MCAR LR High Correlation Variable 10
Means
2.35
2.1
1.85
1.6
1 2 3 4 5 6 7 8 9 10
Repetitions

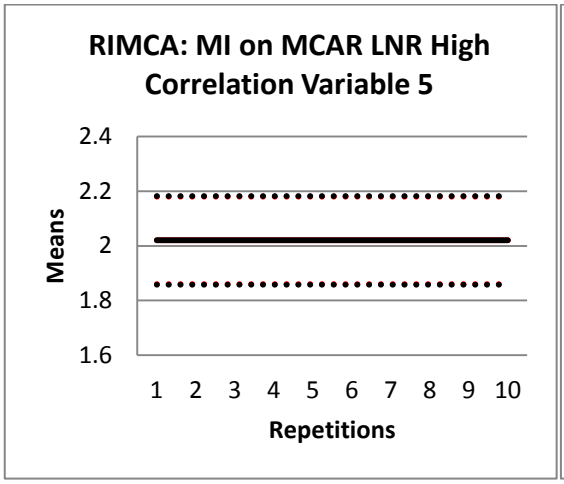
K.8 MCAR non-random pattern with 10% missing values and high correlation structure

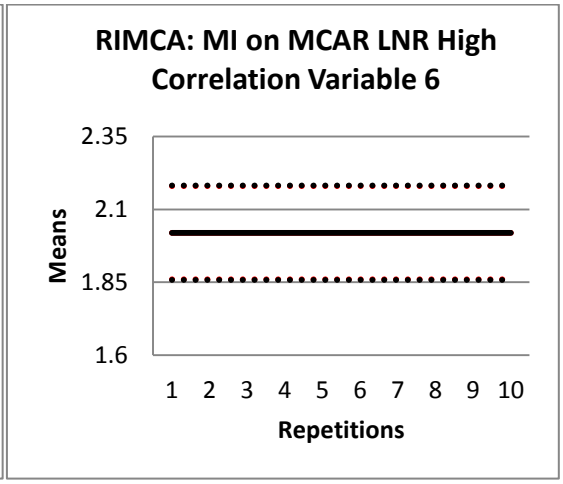


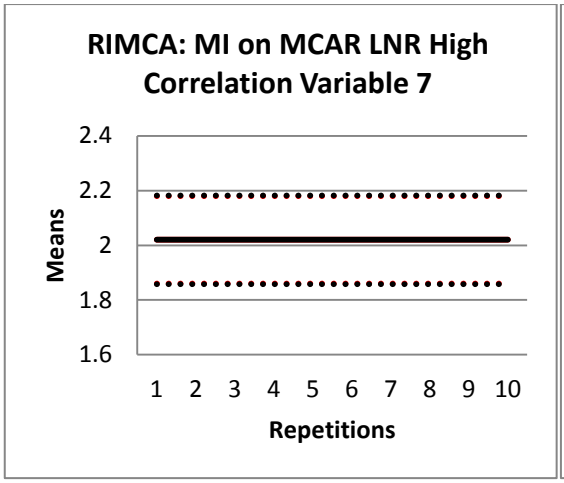


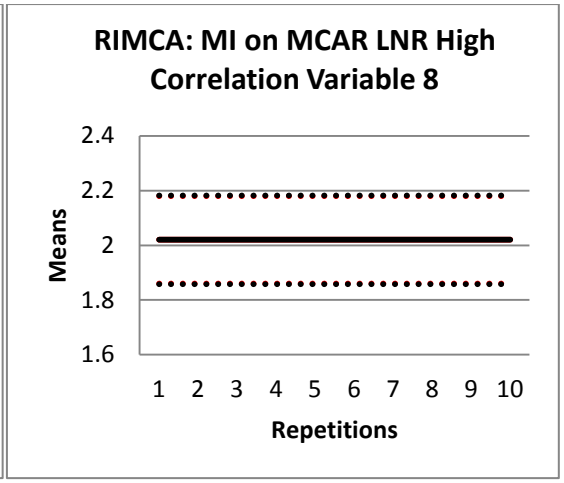


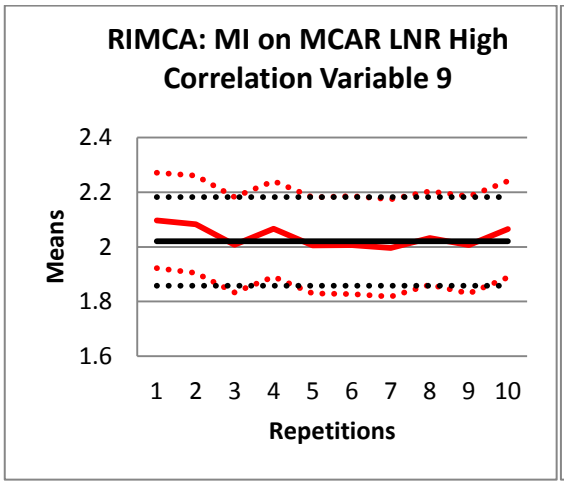


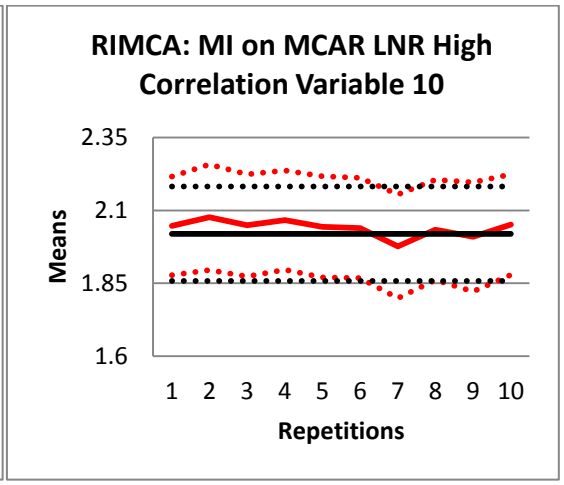




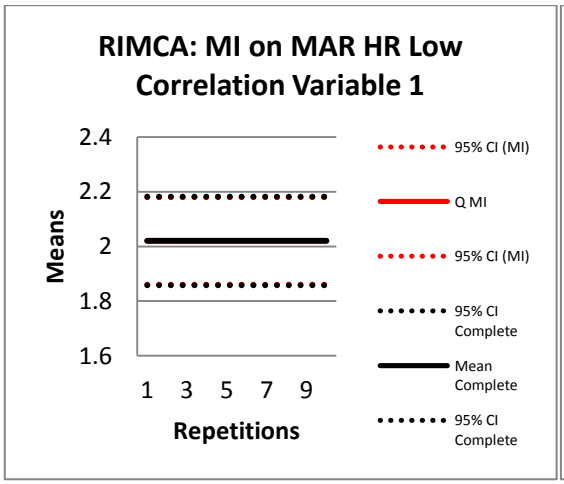


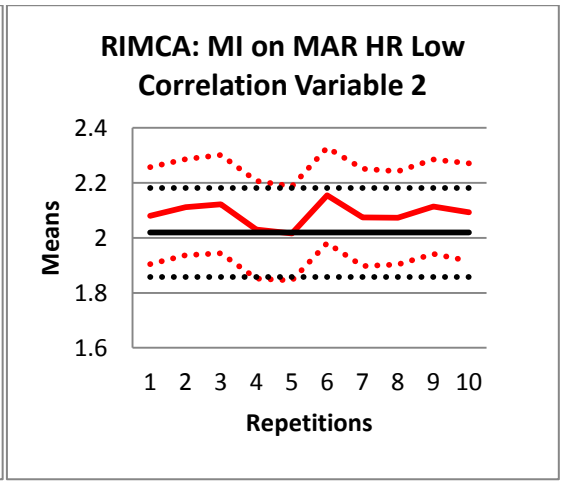


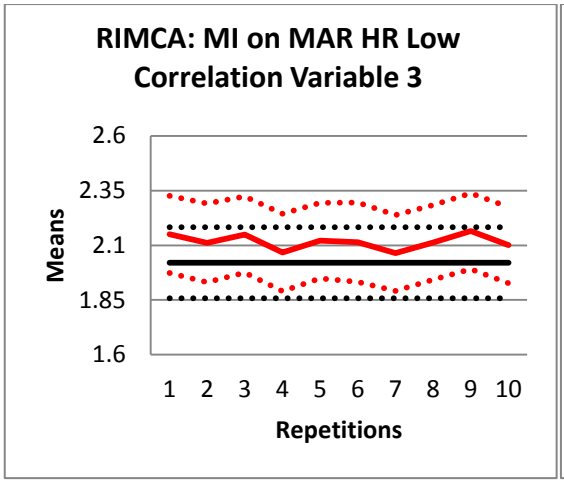


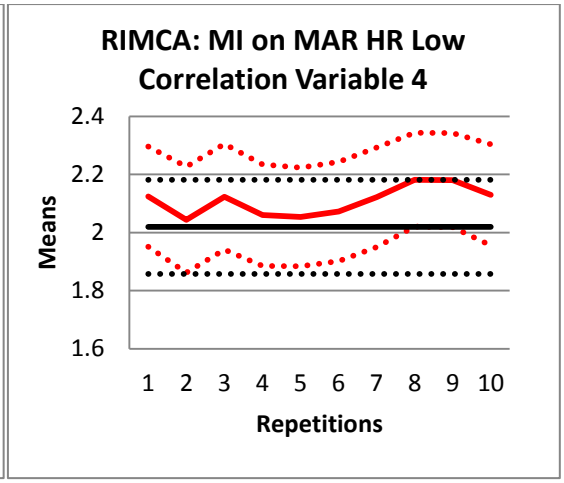


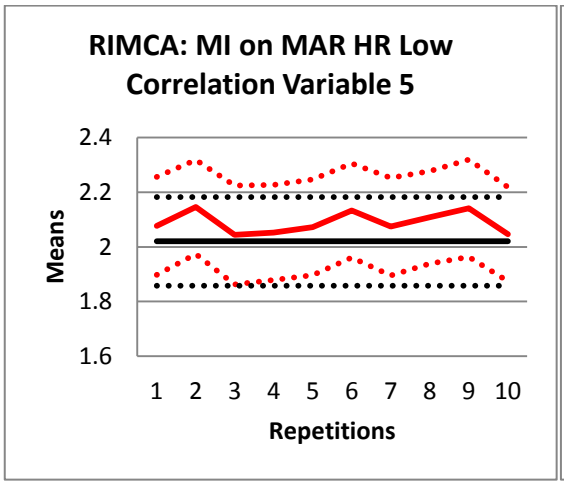
K.9 MAR random pattern with 16% missing values and low correlation structure

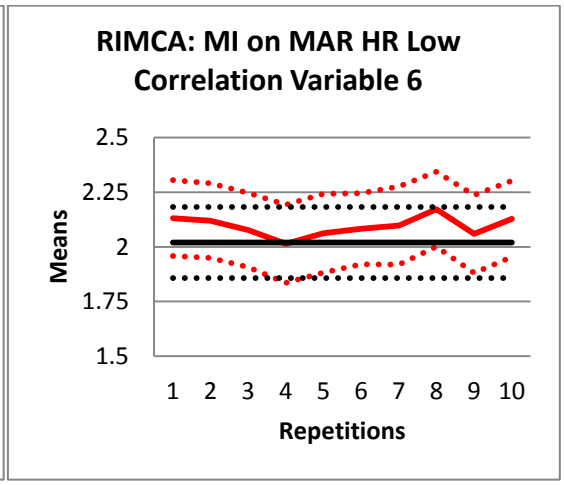


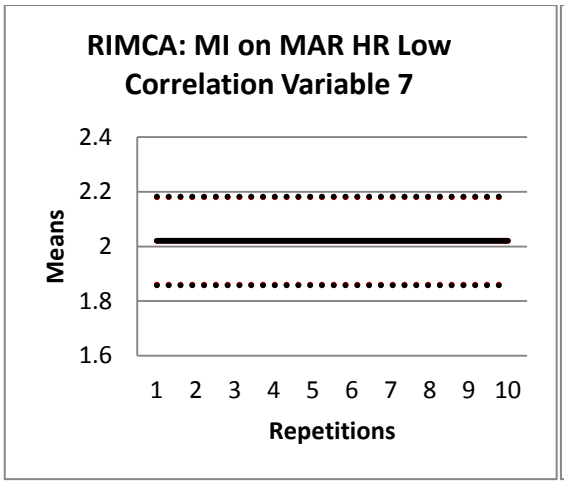


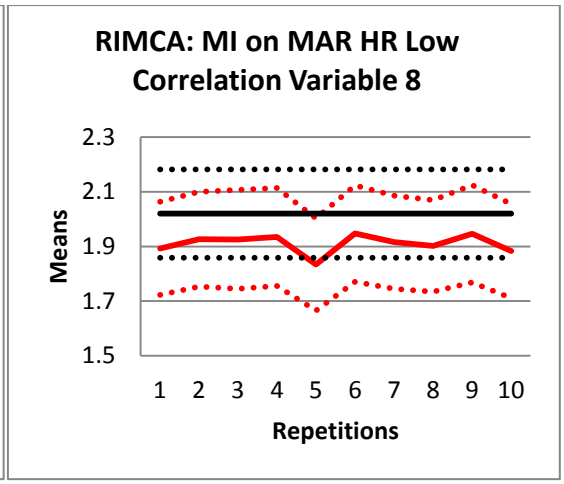

RIMCA: MI on MAR HR Low Correlation Variable 3
Means
2.6
2.35
2.1
1.85
1.6
1 2 3 4 5 6 7 8 9 10
Repetitions

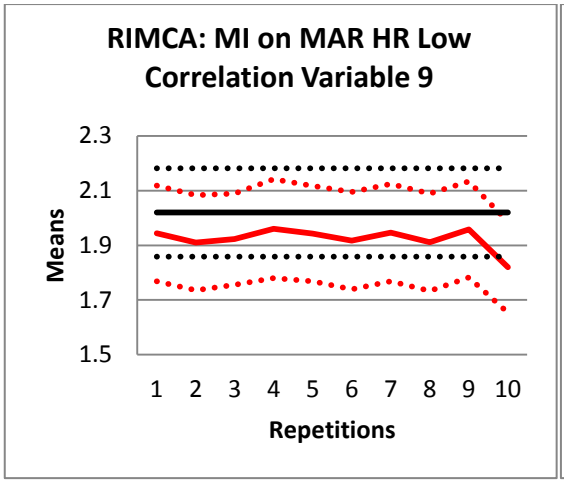

RIMCA: MI on MAR HR Low Correlation Variable 4
Means
2.4
2.2
2
1.8
1.6
1 2 3 4 5 6 7 8 9 10
Repetitions

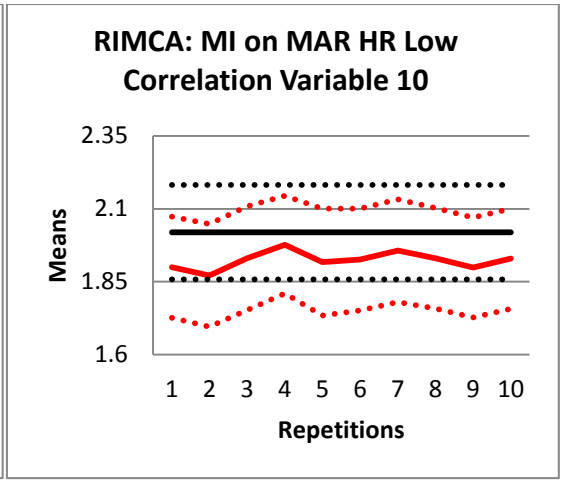

RIMCA: MI on MAR HR Low Correlation Variable 5
Means
2.4
2.2
2
1.8
1.6
1 2 3 4 5 6 7 8 9 10
Repetitions


RIMCA: MI on MAR HR Low Correlation Variable 6
Means
2.5
2.25
2
1.75
1.5
1 2 3 4 5 6 7 8 9 10
Repetitions

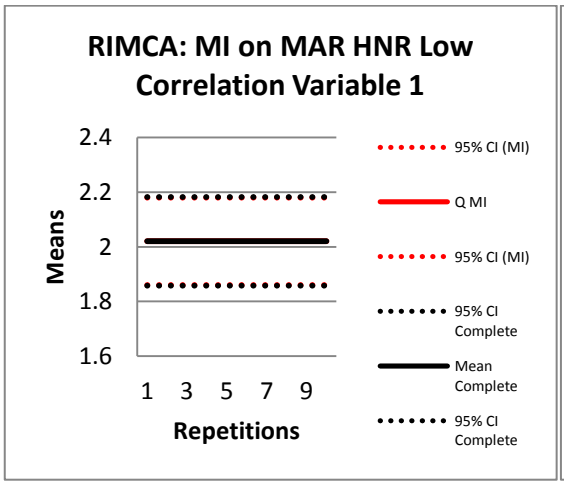

RIMCA: MI on MAR HR Low Correlation Variable 7
Means
2.4
2.2
2
1.8
1.6
1 2 3 4 5 6 7 8 9 10
Repetitions

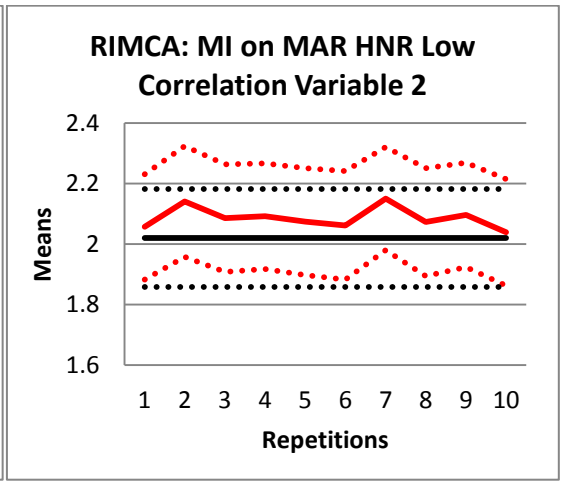

RIMCA: MI on MAR HR Low Correlation Variable 8
Means
2.3
2.1
1.9
1.7
1.5
1 2 3 4 5 6 7 8 9 10
Repetitions

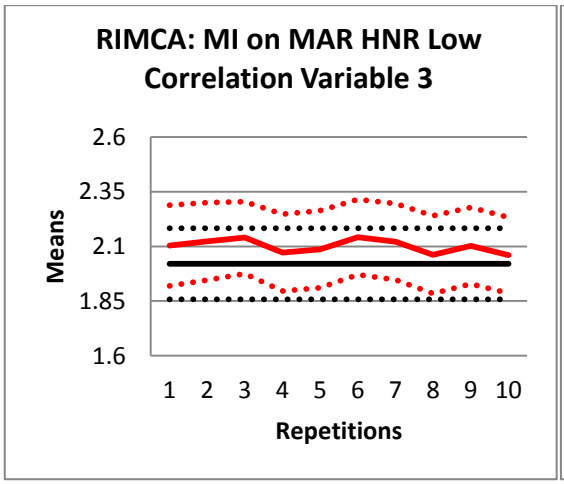


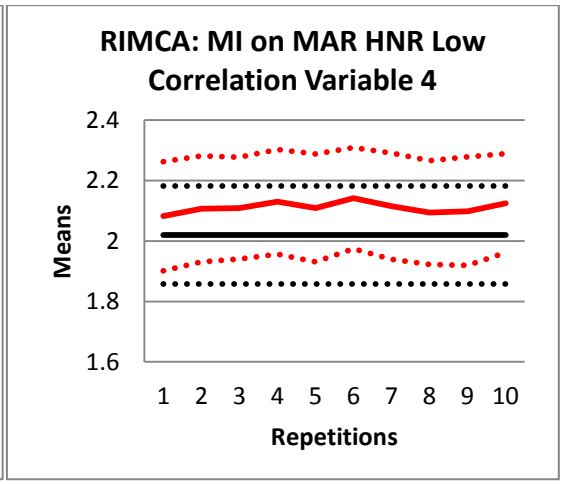


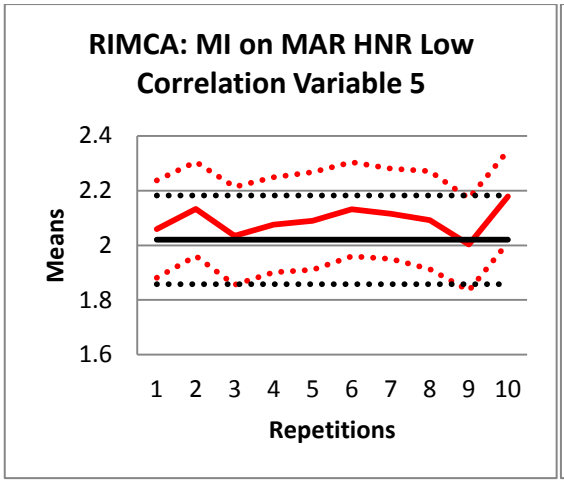
K.10 MAR non-random pattern with 16% missing values and low correlation structure

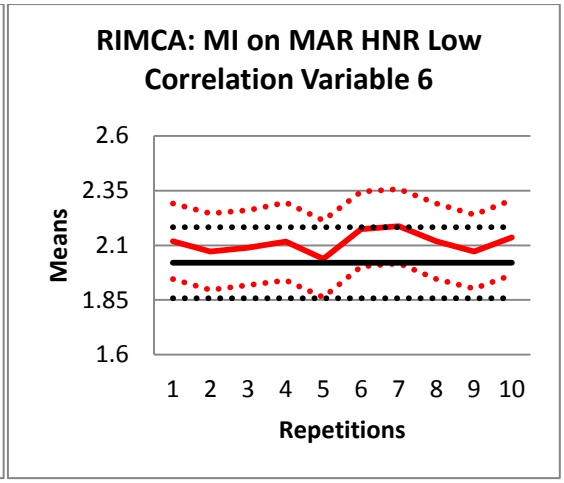


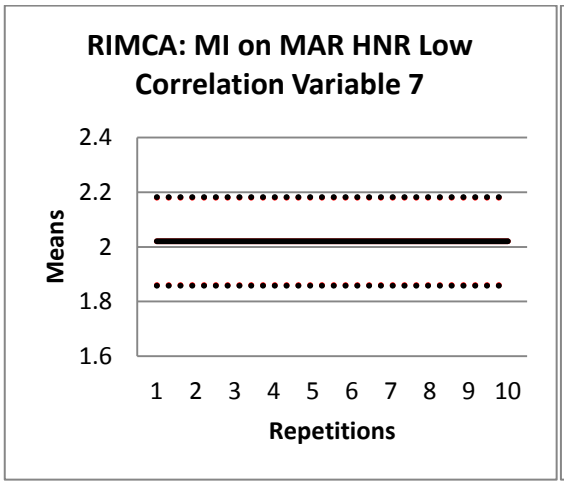


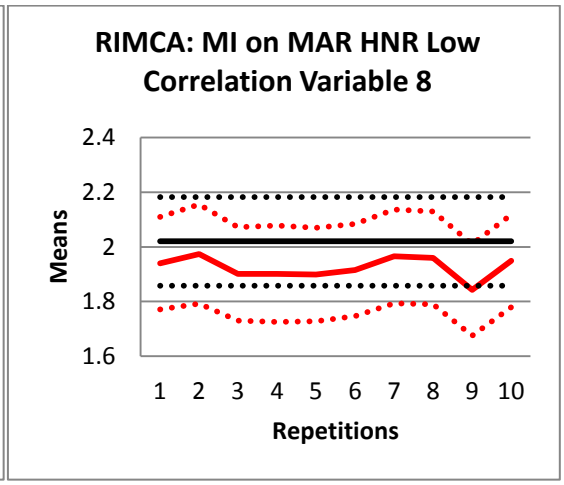


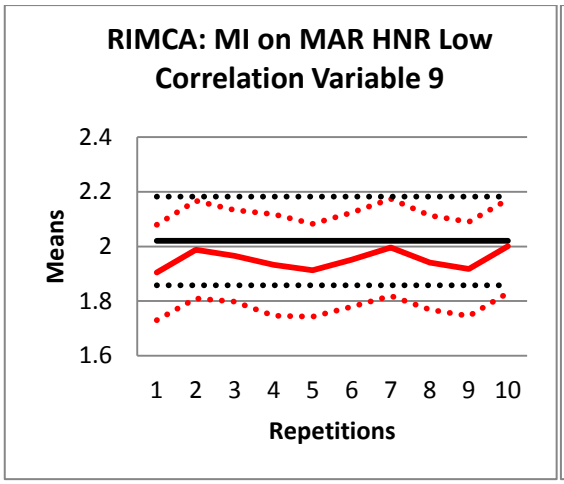


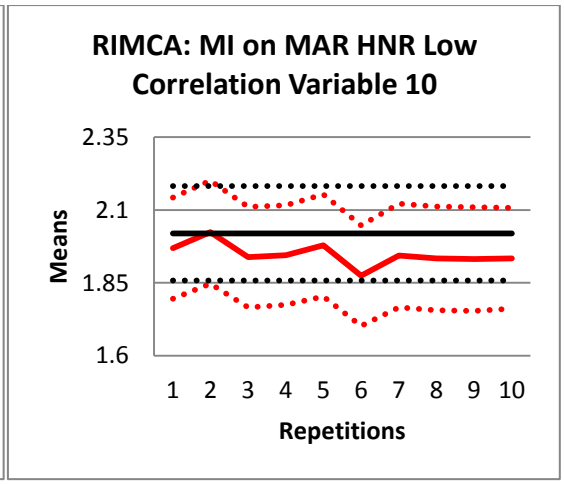

RIMCA: MI on MAR HNR Low Correlation Variable 5
Means
2.4
2.2
2
1.8
1.6
1 2 3 4 5 6 7 8 9 10
Repetitions


RIMCA: MI on MAR HNR Low Correlation Variable 6
Means
2.6
2.35
2.1
1.85
1.6
1 2 3 4 5 6 7 8 9 10
Repetitions

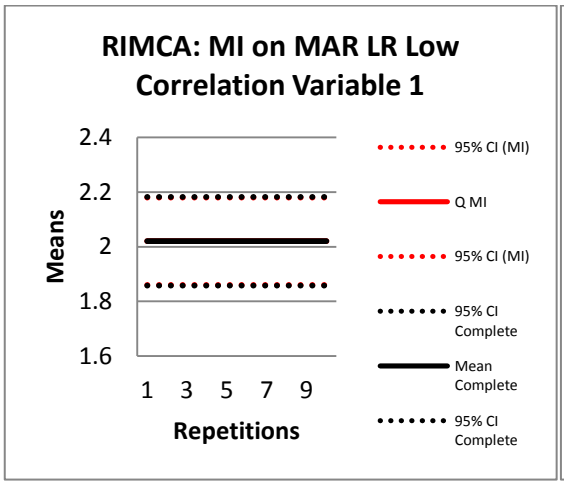

RIMCA: MI on MAR HNR Low Correlation Variable 7
Means
2.4
2.2
2
1.8
1.6
1 2 3 4 5 6 7 8 9 10
Repetitions

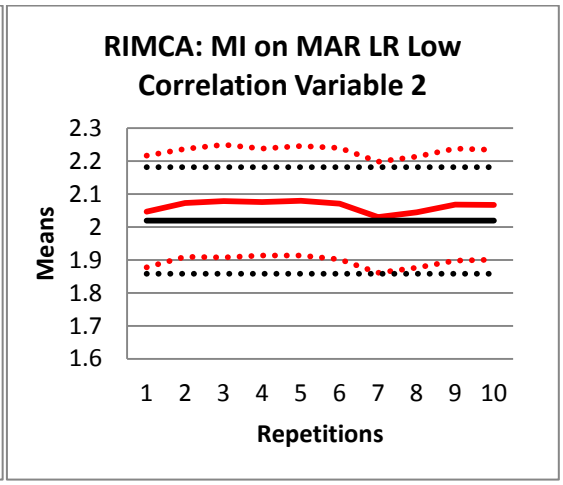

RIMCA: MI on MAR HNR Low Correlation Variable 8
Means
2.4
2.2
2
1.8
1.6
1 2 3 4 5 6 7 8 9 10
Repetitions

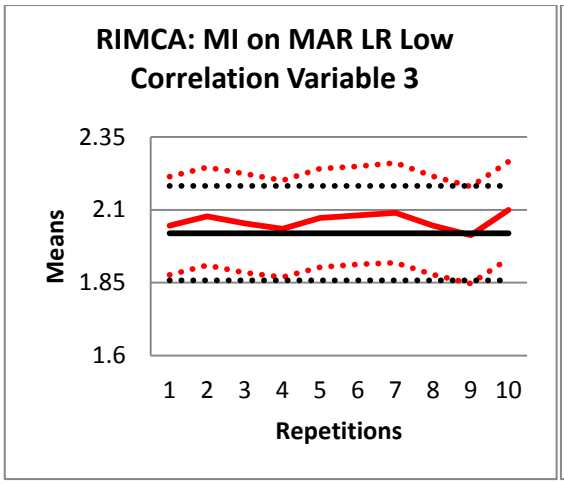

RIMCA: MI on MAR HNR Low Correlation Variable 9
Means
2.4
2.2
2
1.8
1.6
1 2 3 4 5 6 7 8 9 10
Repetitions

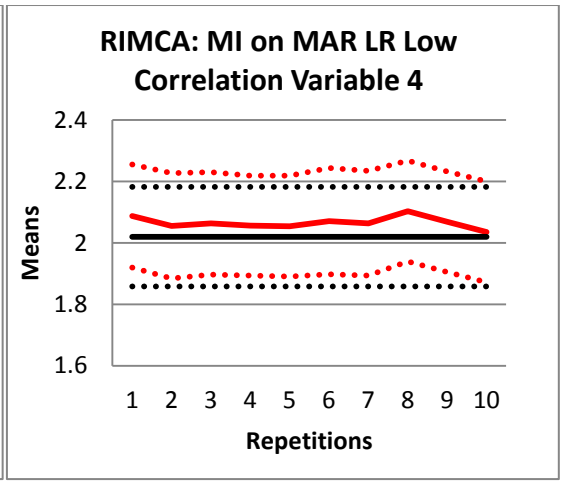

RIMCA: MI on MAR HNR Low Correlation Variable 10
Means
2.35
2.1
1.85
1.6
1 2 3 4 5 6 7 8 9 10
Repetitions

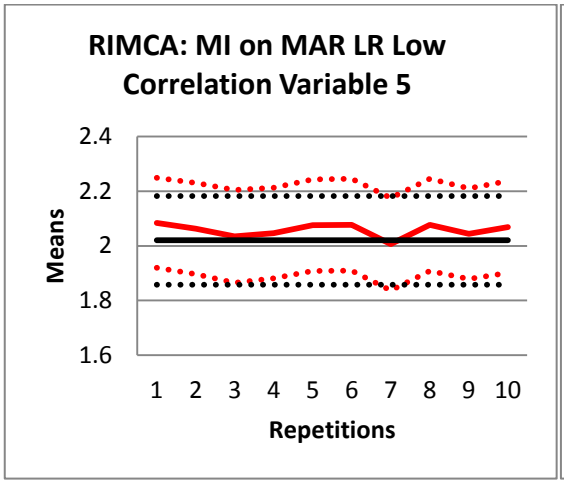
K.11 MAR random pattern with 8% missing values and low correlation structure

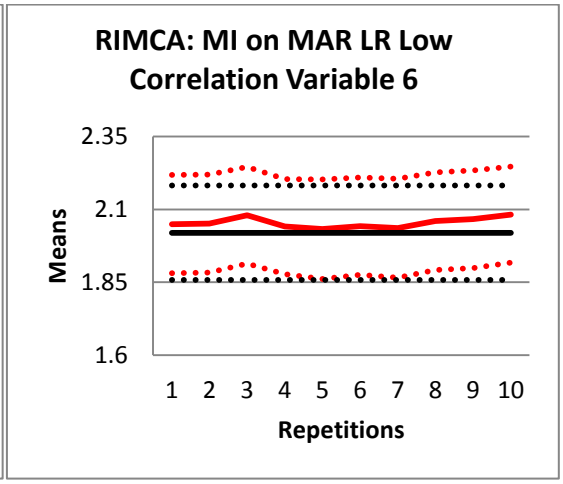


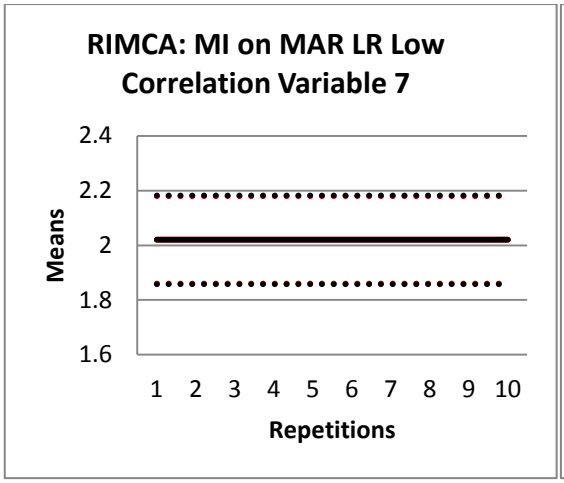


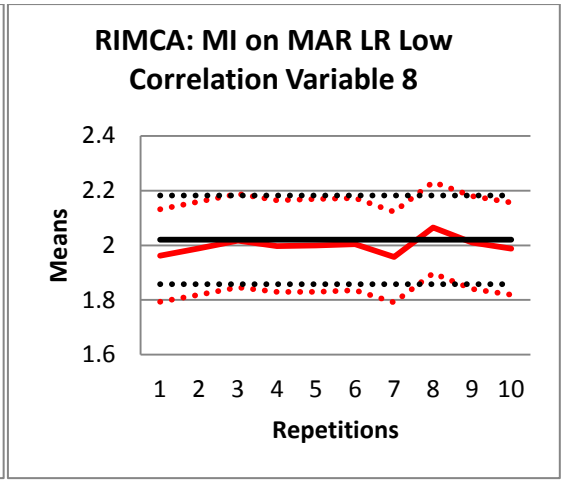


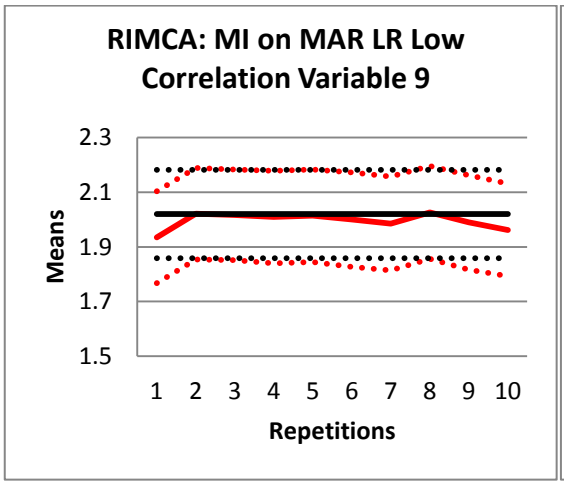


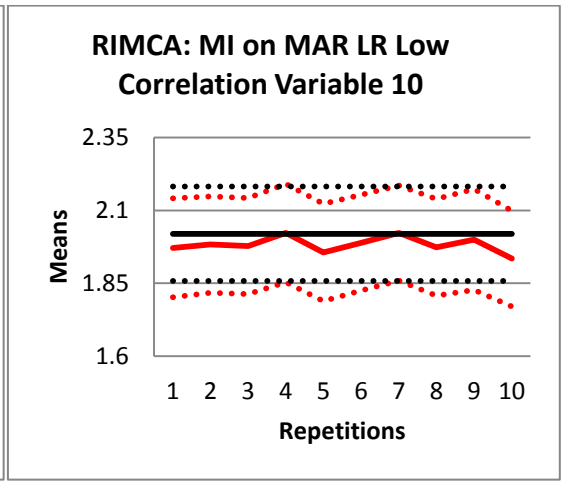




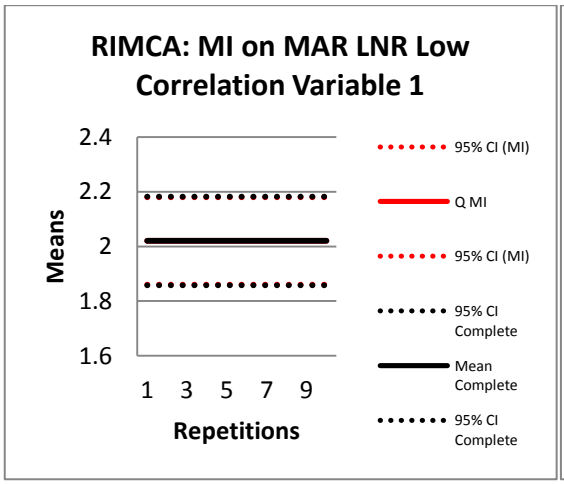


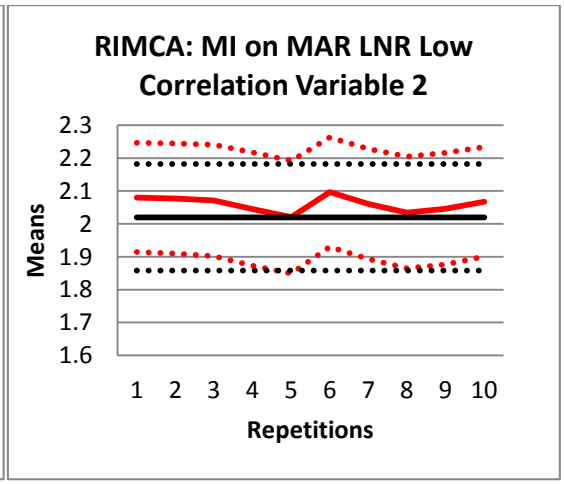


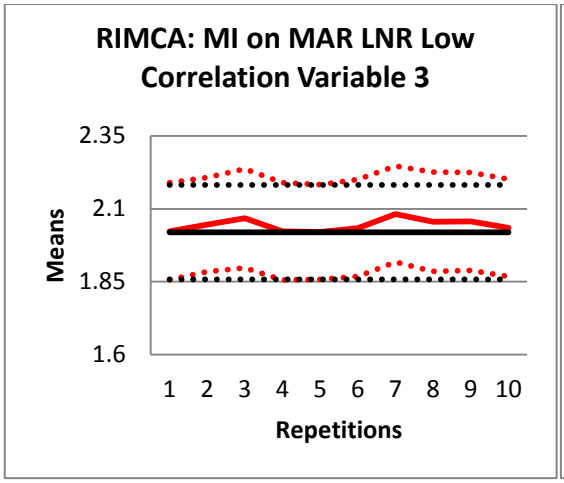


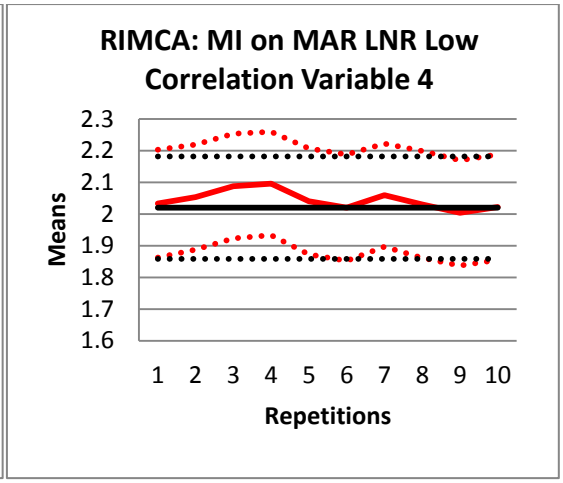


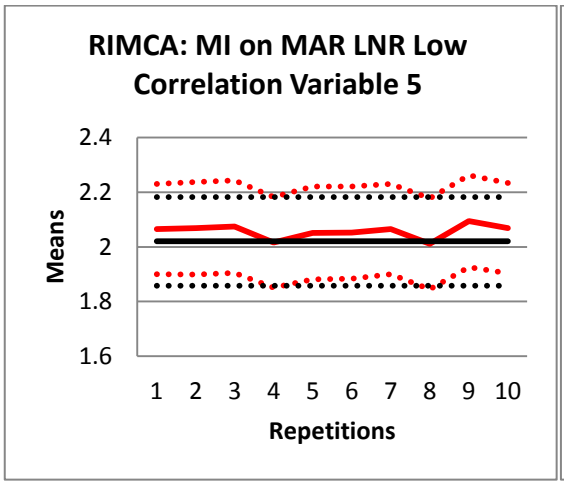
K.12 MAR non-random pattern with 8% missing values and low correlation structure

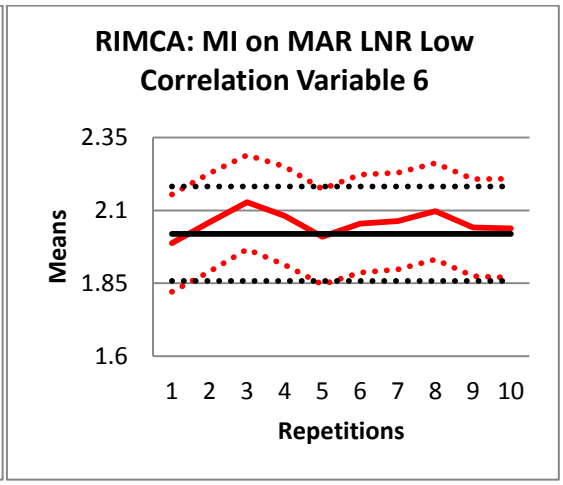


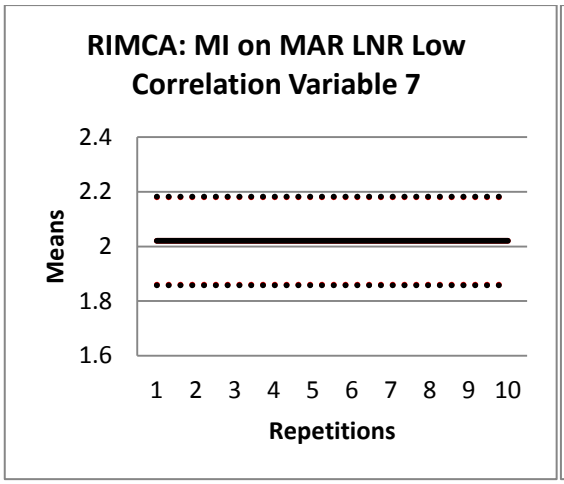


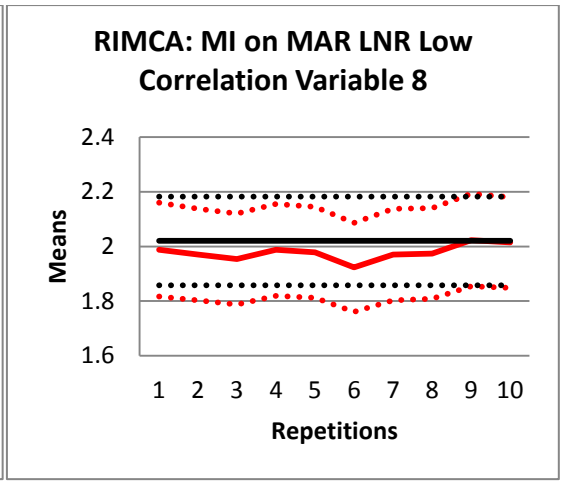

RIMCA: MI on MAR LNR Low Correlation Variable 3
Means
2.35
2.1
1.85
1.6
1 2 3 4 5 6 7 8 9 10
Repetitions

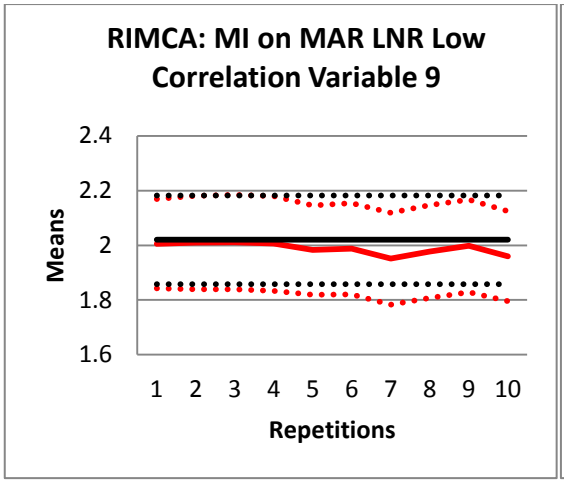

RIMCA: MI on MAR LNR Low Correlation Variable 4
Means
2.3
2.2
2.1
2
1.9
1.8
1.7
1.6
1 2 3 4 5 6 7 8 9 10
Repetitions

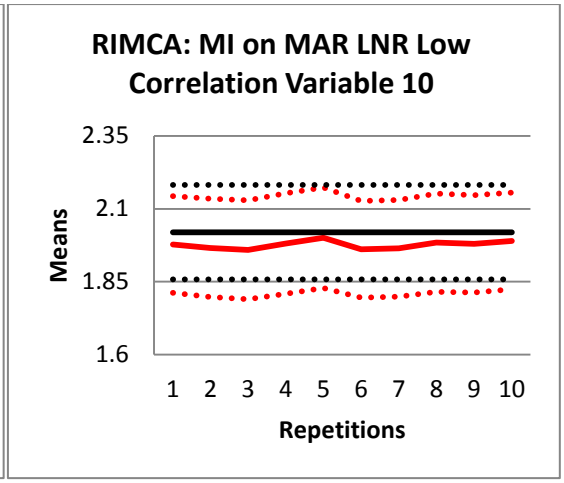

RIMCA: MI on MAR LNR Low Correlation Variable 5
Means
2.4
2.2
2
1.8
1.6
1 2 3 4 5 6 7 8 9 10
Repetitions


RIMCA: MI on MAR LNR Low Correlation Variable 6
Means
2.35
2.1
1.85
1.6
1 2 3 4 5 6 7 8 9 10
Repetitions

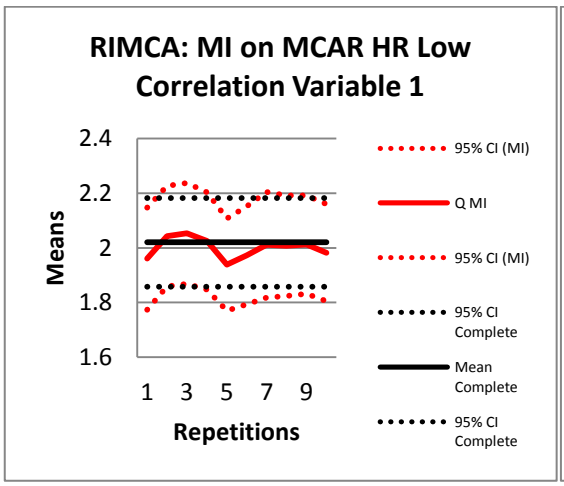

RIMCA: MI on MAR LNR Low Correlation Variable 7
Means
2.4
2.2
2
1.8
1.6
1 2 3 4 5 6 7 8 9 10
Repetitions

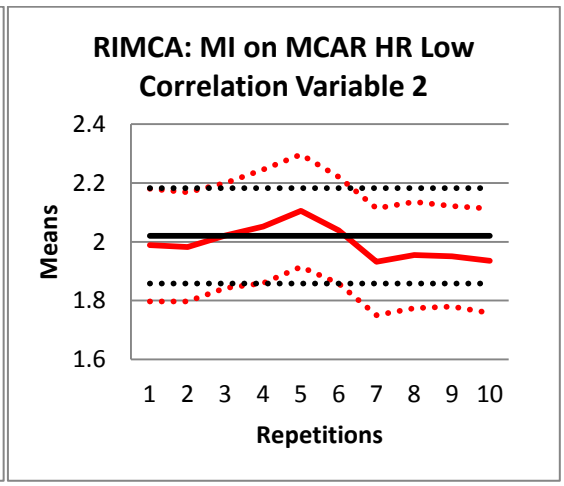

RIMCA: MI on MAR LNR Low Correlation Variable 8
Means
2.4
2.2
2
1.8
1.6
1 2 3 4 5 6 7 8 9 10
Repetitions

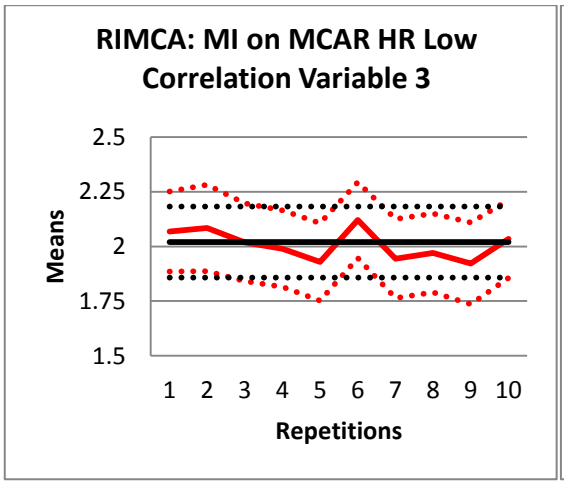


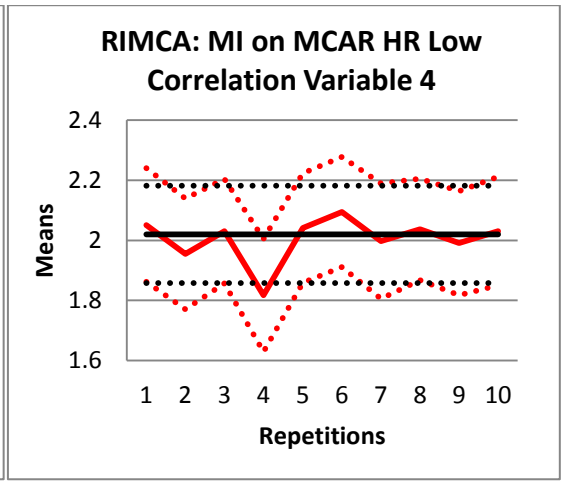


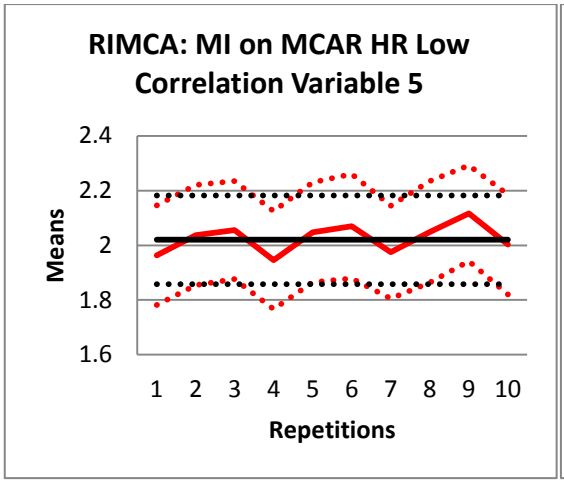
K.13 MCAR random pattern with 30% missing values and low correlation structure

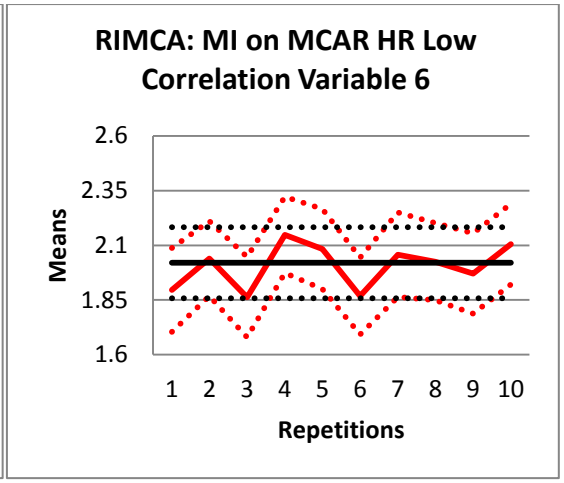


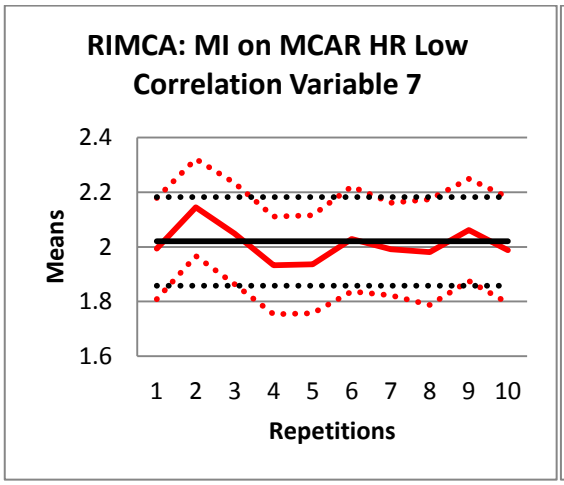


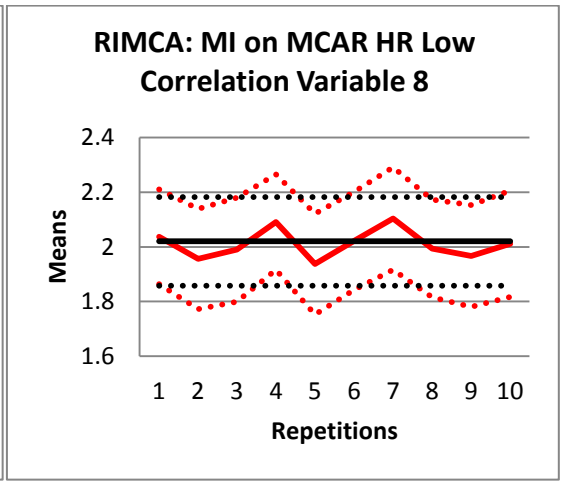


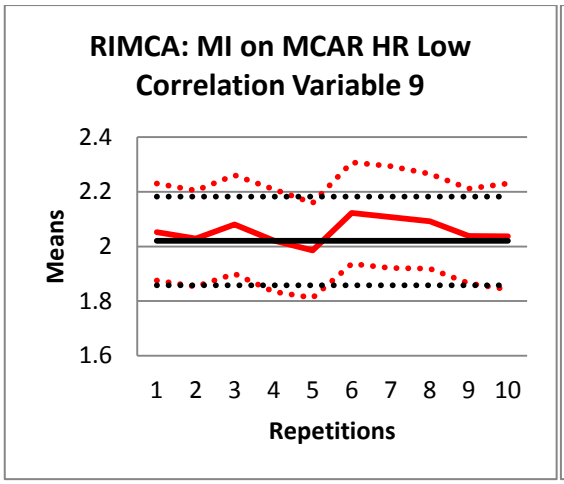


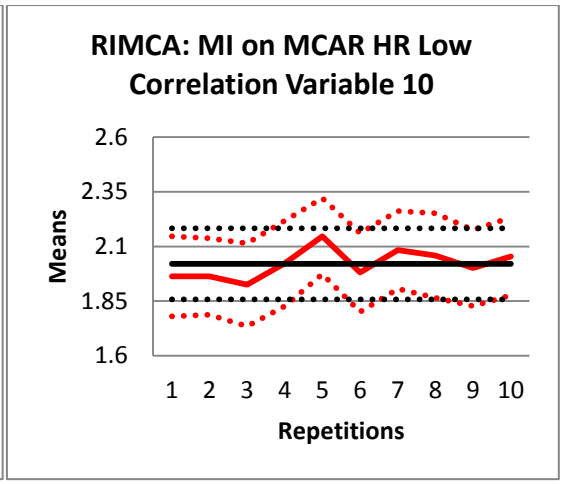

RIMCA: MI on MCAR HR Low Correlation Variable 5
Means
2.4
2.2
2
1.8
1.6
1 2 3 4 5 6 7 8 9 10
Repetitions


RIMCA: MI on MCAR HR Low Correlation Variable 6
Means
2.6
2.35
2.1
1.85
1.6
1 2 3 4 5 6 7 8 9 10
Repetitions

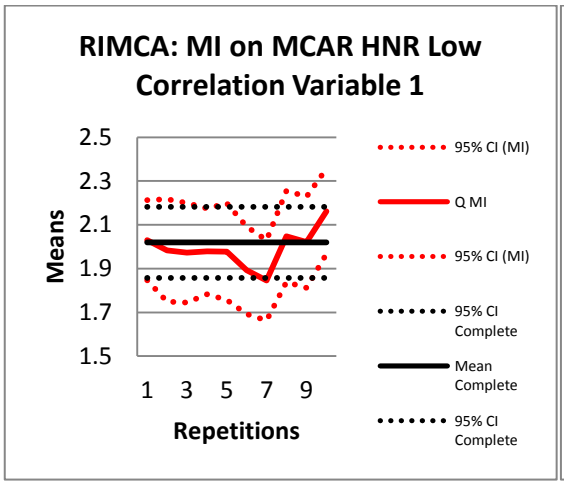

RIMCA: MI on MCAR HR Low Correlation Variable 7
Means
2.4
2.2
2
1.8
1.6
1 2 3 4 5 6 7 8 9 10
Repetitions

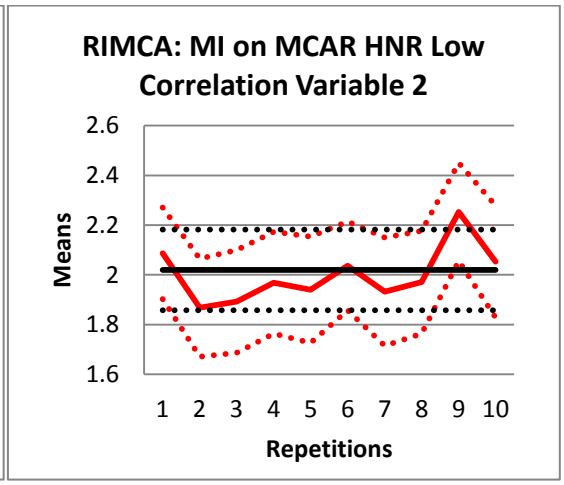

RIMCA: MI on MCAR HR Low Correlation Variable 8
Means
2.4
2.2
2
1.8
1.6
1 2 3 4 5 6 7 8 9 10
Repetitions

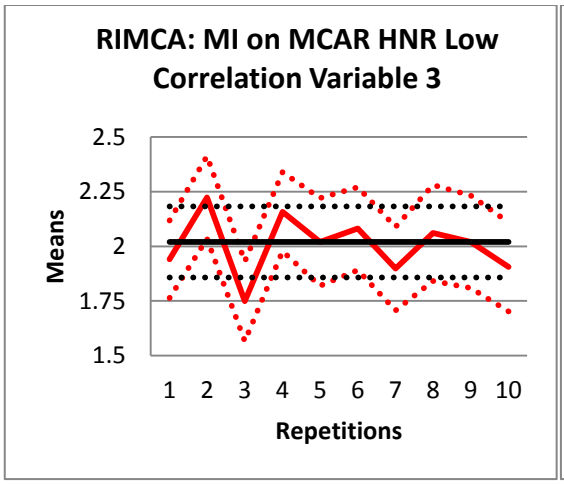

RIMCA: MI on MCAR HR Low Correlation Variable 9
Means
2.4
2.2
2
1.8
1.6
1 2 3 4 5 6 7 8 9 10
Repetitions

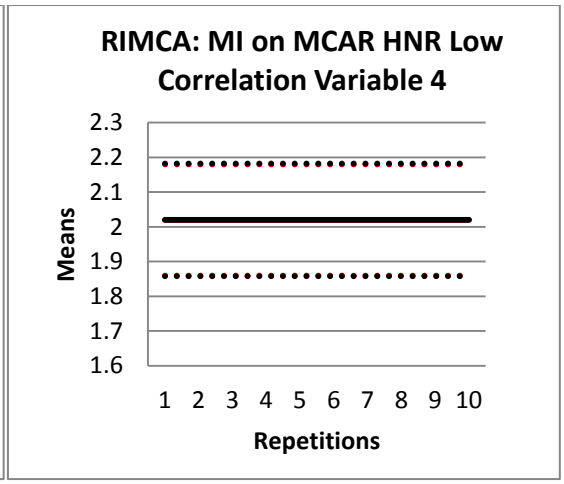

RIMCA: MI on MCAR HR Low Correlation Variable 10
Means
2.6
2.35
2.1
1.85
1.6
1 2 3 4 5 6 7 8 9 10
Repetitions

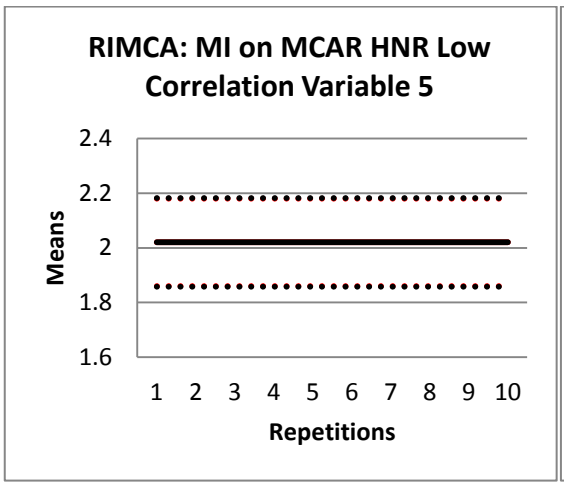
K.14 MCAR non-random pattern with 30% missing values and low correlation structure

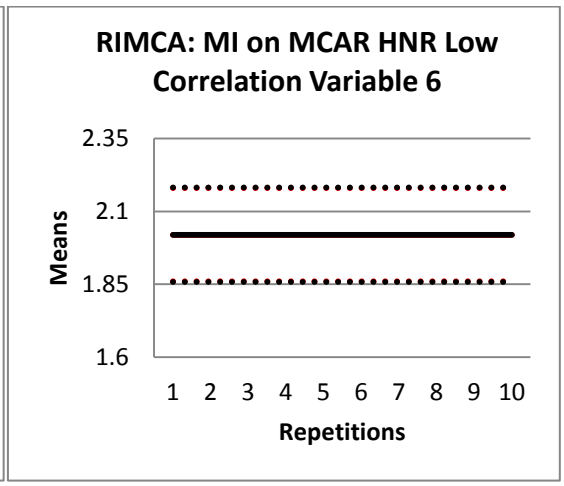


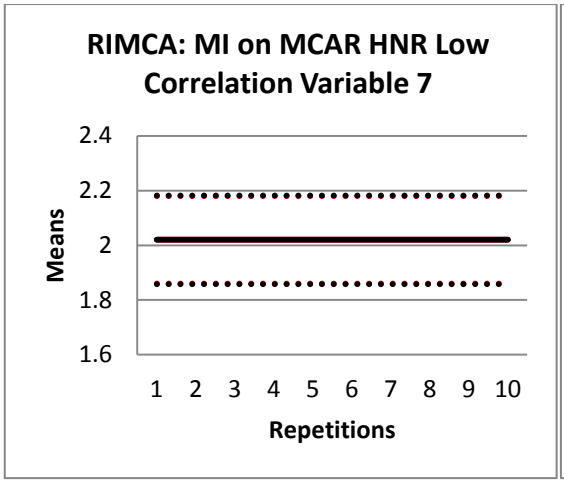


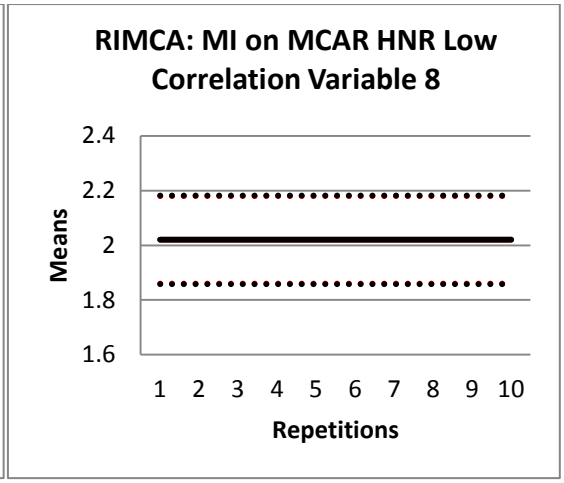


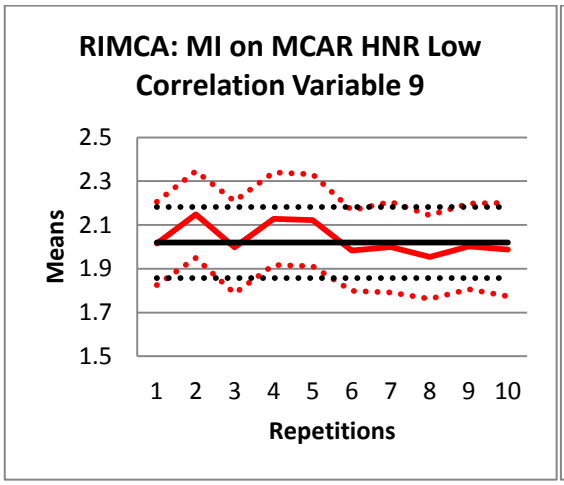


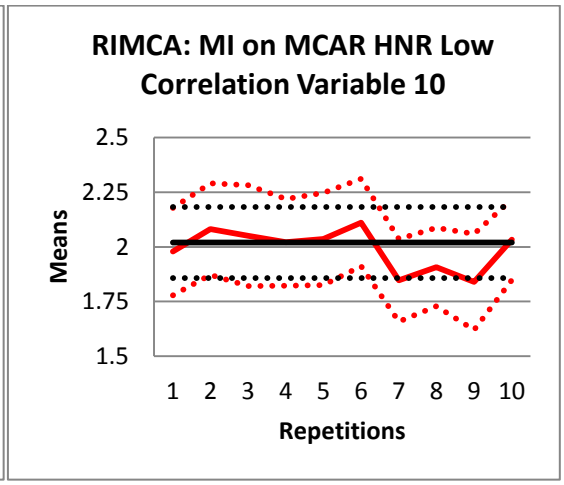




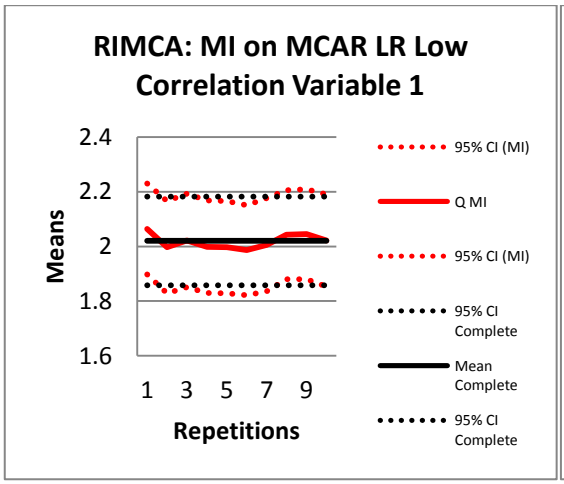


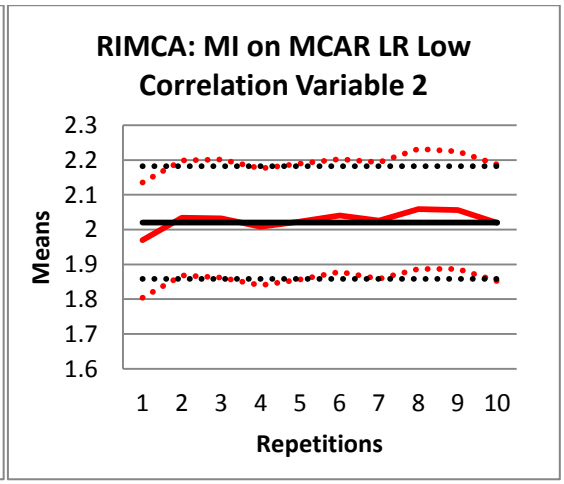


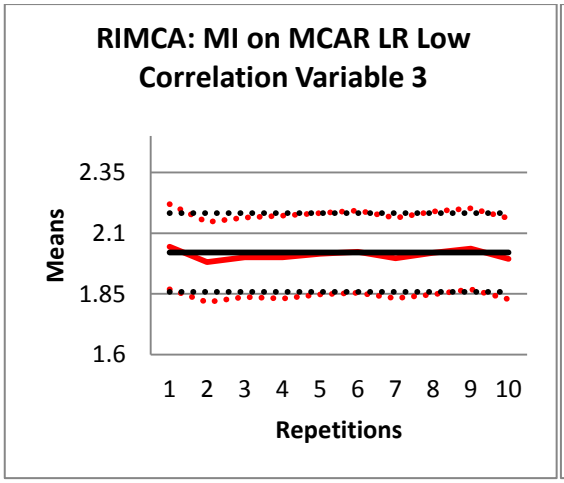


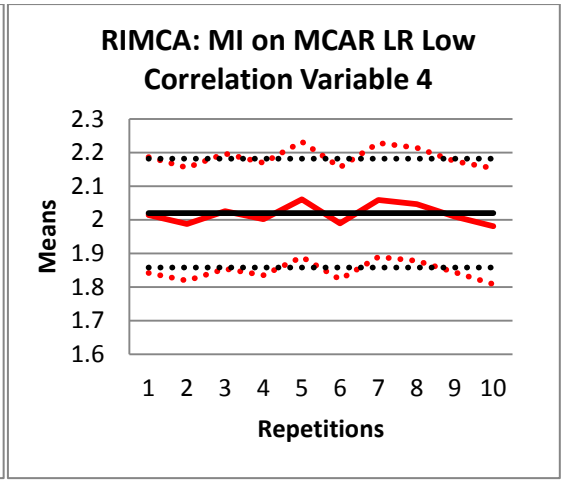


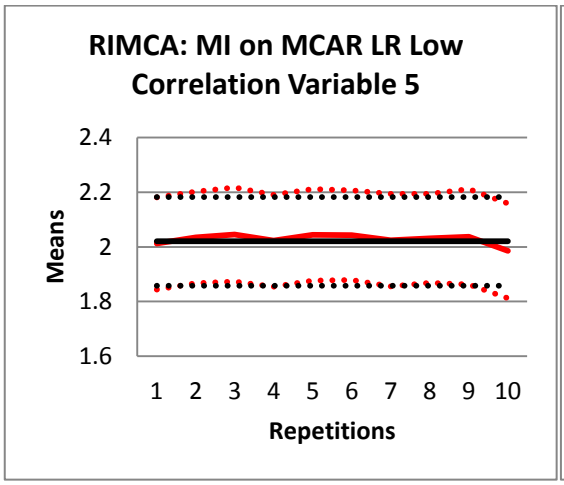
K.15 MCAR random pattern with 10% missing values and low correlation structure

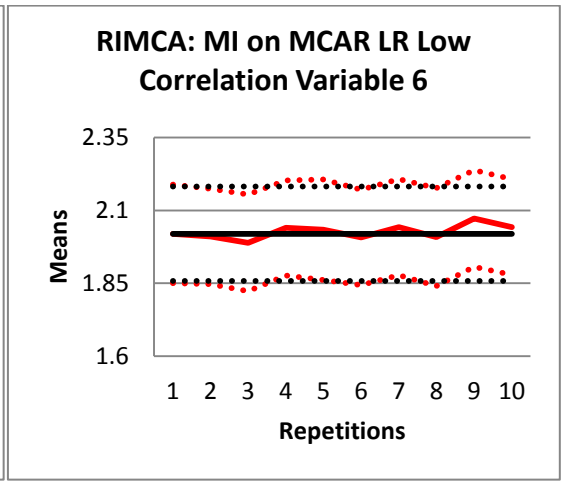


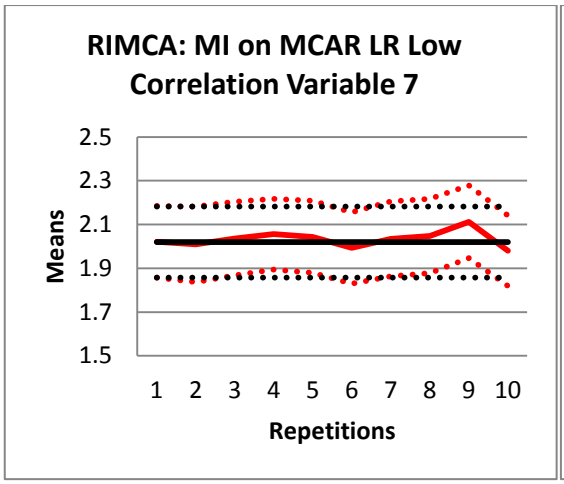


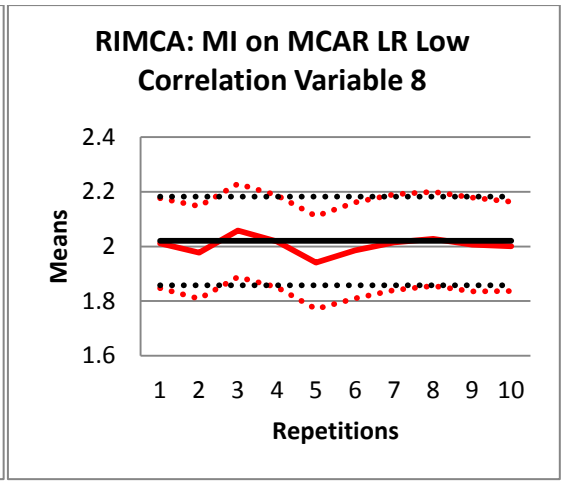

RIMCA: MI on MCAR LR Low Correlation Variable 3
Means
2.35
2.1
1.85
1.6
1 2 3 4 5 6 7 8 9 10
Repetitions

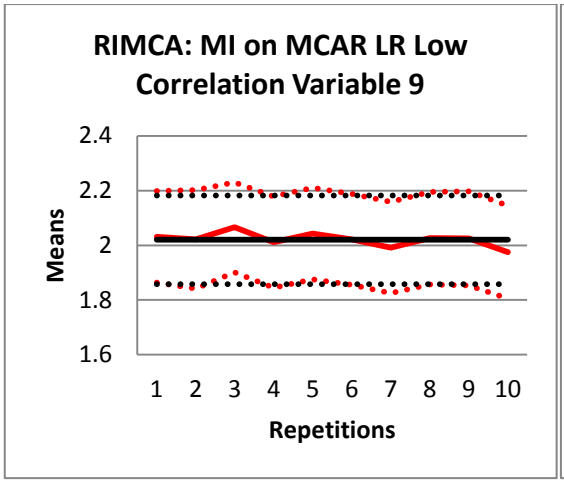

RIMCA: MI on MCAR LR Low Correlation Variable 4
Means
2.3
2.2
2.1
2
1.9
1.8
1.7
1.6
1 2 3 4 5 6 7 8 9 10
Repetitions

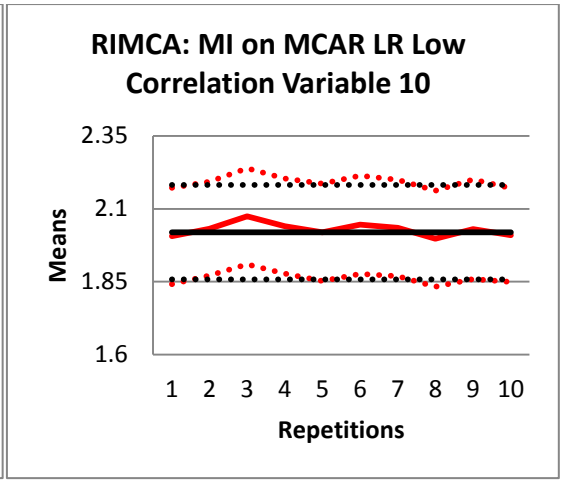

RIMCA: MI on MCAR LR Low Correlation Variable 5
Means
2.4
2.2
2
1.8
1.6
1 2 3 4 5 6 7 8 9 10
Repetitions


RIMCA: MI on MCAR LR Low Correlation Variable 6
Means
2.35
2.1
1.85
1.6
1 2 3 4 5 6 7 8 9 10
Repetitions

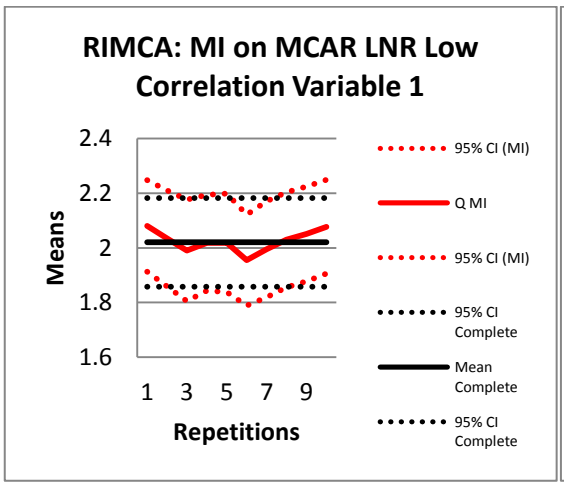

RIMCA: MI on MCAR LR Low Correlation Variable 7
Means
2.5
2.3
2.1
1.9
1.7
1.5
1 2 3 4 5 6 7 8 9 10
Repetitions

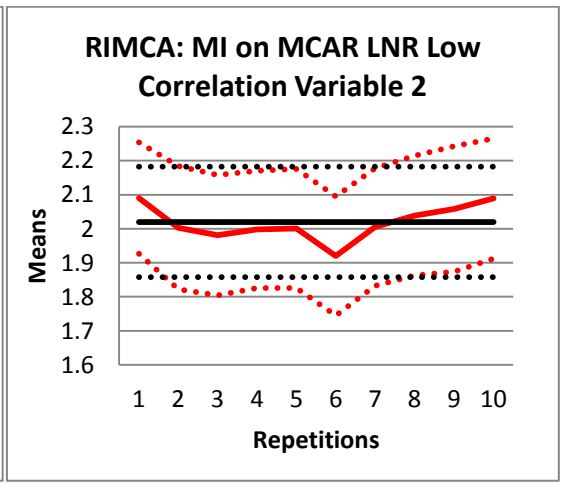

RIMCA: MI on MCAR LR Low Correlation Variable 8
Means
2.4
2.2
2
1.8
1.6
1 2 3 4 5 6 7 8 9 10
Repetitions

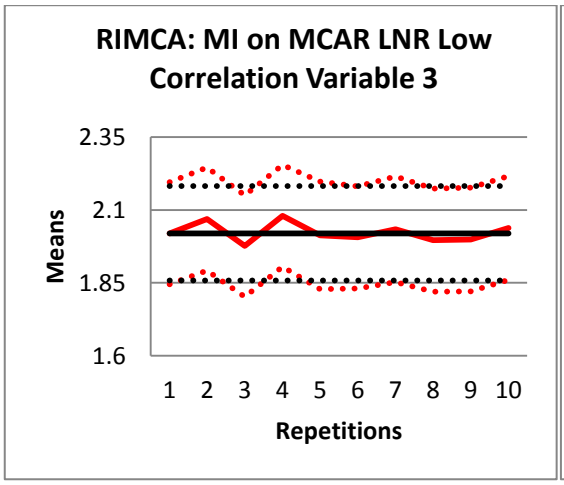


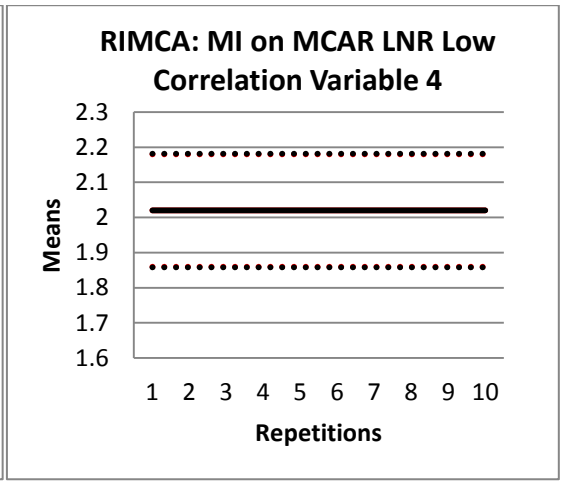


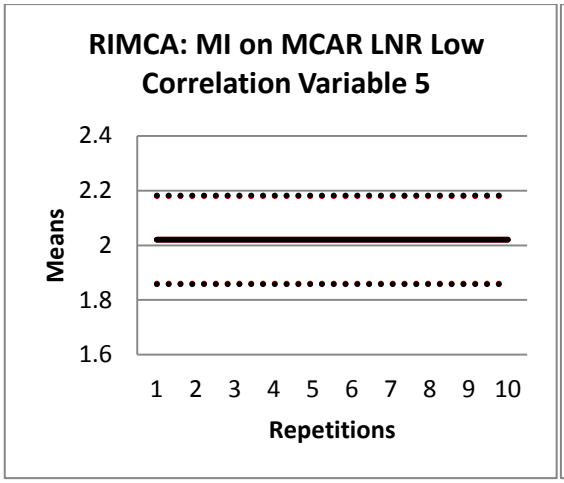
K.16 MCAR non-random pattern with 10% missing values and low correlation structure

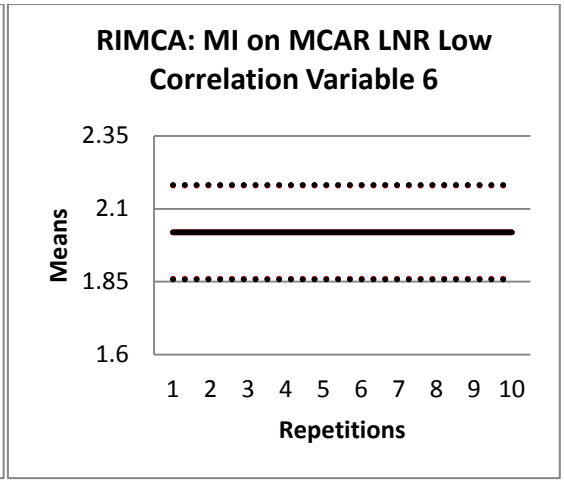


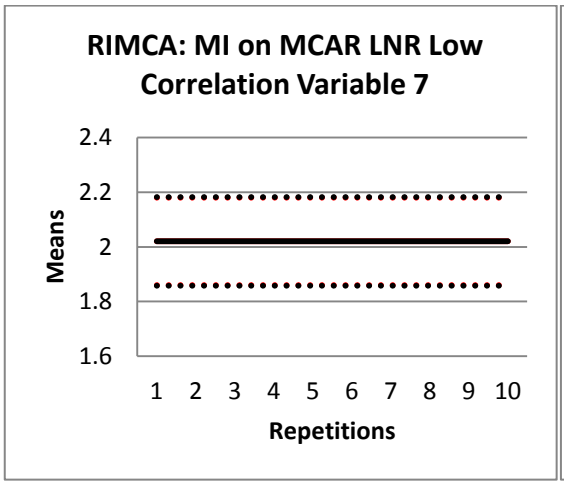


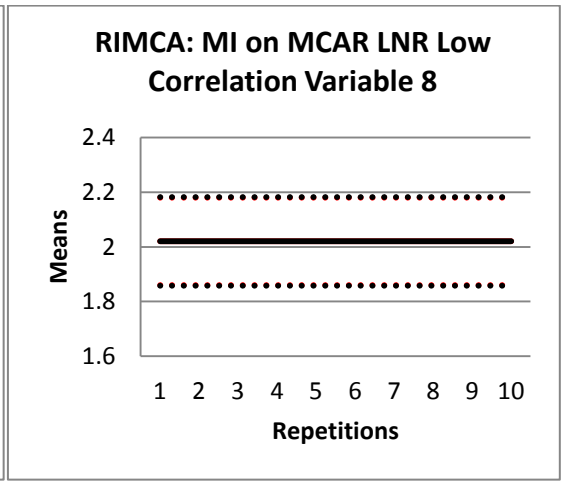


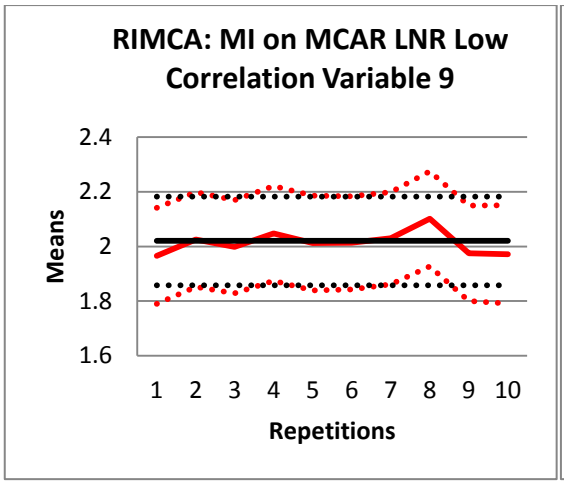


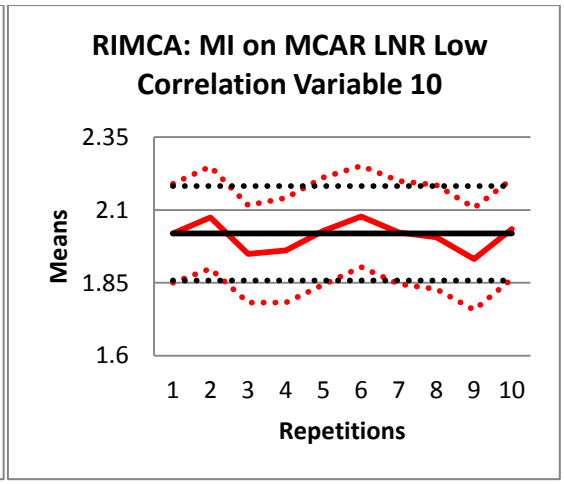

RIMCA: MI on MCAR LNR Low Correlation Variable 5
Means
2.4
2.2
2
1.8
1.6
1 2 3 4 5 6 7 8 9 10
Repetitions


RIMCA: MI on MCAR LNR Low Correlation Variable 6
Means
2.35
2.1
1.85
1.6
1 2 3 4 5 6 7 8 9 10
Repetitions


RIMCA: MI on MCAR LNR Low Correlation Variable 7
Means
2.4
2.2
2
1.8
1.6
1 2 3 4 5 6 7 8 9 10
Repetitions


RIMCA: MI on MCAR LNR Low Correlation Variable 8
Means
2.4
2.2
2
1.8
1.6
1 2 3 4 5 6 7 8 9 10
Repetitions


RIMCA: MI on MCAR LNR Low Correlation Variable 9
Means
2.4
2.2
2
1.8
1.6
1 2 3 4 5 6 7 8 9 10
Repetitions


RIMCA: MI on MCAR LNR Low Correlation Variable 10
Means
2.35
2.1
1.85
1.6
1 2 3 4 5 6 7 8 9 10
Repetitions

Appendix L: Rubin's rules results for simulated data

L.1 MAR HR high and low correlation structure

MAR HR High	Variables									
MI	1	2	3	4	5	6	7	8	9	10
Q bar	2.02	2.22	2.19	2.21	2.20	2.27	2.02	1.92	1.94	1.92
U bar	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
T	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
df	Large	2019.2	1473.1	1374.9	2173.4	1097.6	Large	2580.6	1791.3	2373.2
fmi	0.00	0.16	0.18	0.19	0.15	0.21	0.00	0.14	0.17	0.14
r	0.00	0.18	0.22	0.23	0.18	0.27	0.00	0.16	0.20	0.17
eff	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
t-val	1.96	1.96	1.96	1.96	1.96	1.96	1.96	1.96	1.96	1.96
95% CI	1.86	2.05	2.02	2.05	2.03	2.10	1.86	1.75	1.77	1.74
95% CI	2.18	2.38	2.36	2.38	2.37	2.43	2.18	2.09	2.11	2.10
width	0.32	0.32	0.34	0.33	0.34	0.33	0.32	0.34	0.35	0.35

Q bar –estimated mean for MI (Rubin's rules), U bar –estimated variance for MI (Rubin's rules), B – between-imputation variance, T – total variance, df – degrees of freedom, fmi – fraction of missing information, r – relative increase in variance, eff – efficiency rate, 95% CI – 95% confidence intervals, width – confidence interval width, Large – refers to a very large value which will not be displayed

MAR HR Low	Variables									
MI	1	2	3	4	5	6	7	8	9	10
Q bar	2.02	2.07	2.11	2.02	2.04	2.01	2.02	1.94	1.94	1.96
U bar	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
T	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
df	Large	2099.7	2385.9	1577.3	1860.5	3223.0	Large	2754.5	1494.4	1811.5
fmi	0.00	0.15	0.14	0.18	0.16	0.12	0.00	0.13	0.18	0.17
r	0.00	0.18	0.17	0.21	0.19	0.14	0.00	0.15	0.22	0.20
eff	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
t-val	1.96	1.96	1.96	1.96	1.96	1.96	1.96	1.96	1.96	1.96
95% CI	1.86	1.90	1.94	1.84	1.86	1.84	1.86	1.77	1.77	1.79
95% CI	2.18	2.24	2.28	2.20	2.21	2.18	2.18	2.11	2.12	2.13
width	0.32	0.34	0.34	0.36	0.35	0.34	0.32	0.34	0.35	0.34

Q bar –estimated mean for MI (Rubin's rules), U bar –estimated variance for MI (Rubin's rules), B – between-imputation variance, T – total variance, df – degrees of freedom, fmi – fraction of missing information, r – relative increase in variance, eff – efficiency rate, 95% CI – 95% confidence intervals, width – confidence interval width, Large – refers to a very large value which will not be displayed

L.2 MAR HNR high and low correlation structure

MAR HNR High	Variables									
MI	1	2	3	4	5	6	7	8	9	10
Q bar	2.02	2.18	2.18	2.22	2.20	2.25	2.02	1.85	1.86	1.87
U bar	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
T	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
df	Large	2047.4	845.87	1045.6	1664.3	1983.8	Large	1816.8	2299.1	2864.5
fmi	0.00	0.16	0.24	0.22	0.17	0.16	0.00	0.17	0.15	0.13
r	0.00	0.18	0.32	0.28	0.21	0.19	0.00	0.20	0.17	0.15
eff	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
t-val	1.96	1.96	1.96	1.96	1.96	1.96	1.96	1.96	1.96	1.96
95% CI	1.86	2.01	2.00	2.05	2.04	2.09	1.86	1.69	1.69	1.70
95% CI	2.18	2.34	2.36	2.38	2.37	2.41	2.18	2.02	2.02	2.03
width	0.32	0.33	0.36	0.33	0.33	0.31	0.32	0.34	0.33	0.34

Q bar –estimated mean for MI (Rubin's rules), U bar –estimated variance for MI (Rubin's rules), B – between-imputation variance, T – total variance, df – degrees of freedom, fmi – fraction of missing information, r – relative increase in variance, eff – efficiency rate, 95% CI – 95% confidence intervals, width – confidence interval width, Large – refers to a very large value which will not be displayed

MAR HNR Low	Variables									
MI	1	2	3	4	5	6	7	8	9	10
Q bar	2.02	2.07	2.13	2.06	2.06	2.04	2.02	1.97	1.91	1.91
U bar	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
T	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
df	Large	2532.5	1989.9	1222.5	1496.5	1188.2	Large	2506.0	1819.1	1419.0
fmi	0.00	0.14	0.16	0.20	0.18	0.20	0.00	0.14	0.17	0.19
r	0.00	0.16	0.19	0.25	0.22	0.25	0.00	0.16	0.20	0.23
eff	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
t-val	1.96	1.96	1.96	1.96	1.96	1.96	1.96	1.96	1.96	1.96
95% CI	1.86	1.90	1.96	1.87	1.89	1.86	1.86	1.80	1.74	1.74
95% CI	2.18	2.25	2.30	2.24	2.23	2.22	2.18	2.15	2.08	2.09
width	0.32	0.35	0.35	0.37	0.35	0.36	0.32	0.35	0.34	0.35

Q bar –estimated mean for MI (Rubin's rules), U bar –estimated variance for MI (Rubin's rules), B – between-imputation variance, T – total variance, df – degrees of freedom, fmi – fraction of missing information, r – relative increase in variance, eff – efficiency rate, 95% CI – 95% confidence intervals, width – confidence interval width, Large – refers to a very large value which will not be displayed

L.3 MAR LR high and low correlation structure

MAR LR High	Variables									
MI	1	2	3	4	5	6	7	8	9	10
Q bar	2.02	2.05	2.09	2.09	2.08	2.11	2.02	1.92	1.93	1.93
U bar	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
T	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
df	Large	9366.0	5264	8238.1	5251.2	5496.1	Large	9089.7	12978	8860.7
fmi	0.00	0.07	0.10	0.08	0.10	0.09	0.00	0.07	0.06	0.07
r	0.00	0.08	0.11	0.08	0.11	0.10	0.00	0.08	0.07	0.08
eff	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
t-val	1.96	1.96	1.96	1.96	1.96	1.96	1.96	1.96	1.96	1.96
95% CI	1.86	1.89	1.93	1.93	1.92	1.95	1.86	1.75	1.76	1.76
95% CI	2.18	2.22	2.26	2.26	2.25	2.27	2.18	2.08	2.09	2.09
width	0.32	0.33	0.33	0.33	0.33	0.33	0.32	0.32	0.32	0.33

Q bar –estimated mean for MI (Rubin's rules), U bar –estimated variance for MI (Rubin's rules), B – between-imputation variance, T – total variance, df – degrees of freedom, fmi – fraction of missing information, r – relative increase in variance, eff – efficiency rate, 95% CI – 95% confidence intervals, width – confidence interval width, Large – refers to a very large value which will not be displayed

MAR LR Low	Variables									
MI	1	2	3	4	5	6	7	8	9	10
Q bar	2.02	2.00	2.02	2.00	2.05	2.02	2.02	2.03	1.99	1.94
U bar	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
T	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
df	Large	10490	6864.6	7658.6	5491.3	6022.0	Large	5890.0	8421.0	4973.6
fmi	0.00	0.07	0.08	0.08	0.09	0.09	0.00	0.09	0.08	0.10
r	0.00	0.07	0.09	0.09	0.10	0.10	0.00	0.10	0.08	0.11
eff	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
t-val	1.96	1.96	1.96	1.96	1.96	1.96	1.96	1.96	1.96	1.96
95% CI	1.86	1.83	1.85	1.83	1.89	1.85	1.86	1.86	1.83	1.77
95% CI	2.18	2.16	2.19	2.17	2.22	2.19	2.18	2.20	2.16	2.11
width	0.32	0.33	0.34	0.34	0.34	0.34	0.32	0.34	0.33	0.33

Q bar –estimated mean for MI (Rubin's rules), U bar –estimated variance for MI (Rubin's rules), B – between-imputation variance, T – total variance, df – degrees of freedom, fmi – fraction of missing information, r – relative increase in variance, eff – efficiency rate, 95% CI – 95% confidence intervals, width – confidence interval width, Large – refers to a very large value which will not be displayed

L.4 MAR LNR high and low correlation structure

MAR LNR High	Variables									
MI	1	2	3	4	5	6	7	8	9	10
Q bar	2.02	2.10	2.09	2.12	2.10	2.13	2.02	1.98	1.98	1.96
U bar	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
T	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
df	Large	6165	4235.7	4173.6	4348.2	6123.4	Large	2821.4	5756.5	6199.4
fmi	0.00	0.09	0.11	0.11	0.11	0.09	0.00	0.13	0.09	0.09
r	0.00	0.10	0.12	0.12	0.12	0.10	0.00	0.15	0.10	0.10
eff	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
t-val	1.96	1.96	1.96	1.96	1.96	1.96	1.96	1.96	1.96	1.96
95% CI	1.86	1.94	1.92	1.96	1.93	1.97	1.86	1.80	1.81	1.79
95% CI	2.18	2.26	2.26	2.29	2.26	2.29	2.18	2.15	2.14	2.13
width	0.32	0.33	0.34	0.33	0.34	0.32	0.32	0.35	0.33	0.34

Q bar –estimated mean for MI (Rubin's rules), U bar –estimated variance for MI (Rubin's rules), B – between-imputation variance, T – total variance, df – degrees of freedom, fmi – fraction of missing information, r – relative increase in variance, eff – efficiency rate, 95% CI – 95% confidence intervals, width – confidence interval width, Large – refers to a very large value which will not be displayed

MAR LNR Low	Variables									
MI	1	2	3	4	5	6	7	8	9	10
Q bar	2.02	2.06	2.06	2.04	2.04	2.05	2.02	1.99	1.97	1.98
U bar	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
T	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
df	Large	3782.7	5450.4	5182	5324.1	6965.7	Large	11024	4900.2	16417
fmi	0.00	0.11	0.10	0.10	0.10	0.08	0.00	0.07	0.10	0.05
r	0.00	0.13	0.10	0.11	0.11	0.09	0.00	0.07	0.11	0.06
eff	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
t-val	1.96	1.96	1.96	1.96	1.96	1.96	1.96	1.96	1.96	1.96
95% CI	1.86	1.89	1.90	1.87	1.88	1.88	1.86	1.83	1.81	1.82
95% CI	2.18	2.23	2.23	2.21	2.21	2.21	2.18	2.15	2.14	2.14
width	0.32	0.34	0.34	0.34	0.33	0.34	0.32	0.32	0.33	0.32

Q bar –estimated mean for MI (Rubin's rules), U bar –estimated variance for MI (Rubin's rules), B – between-imputation variance, T – total variance, df – degrees of freedom, fmi – fraction of missing information, r – relative increase in variance, eff – efficiency rate, 95% CI – 95% confidence intervals, width – confidence interval width, Large – refers to a very large value which will not be displayed

L.5 MCAR HR high and low correlation structure

MCAR HR High	Variables									
MI	1	2	3	4	5	6	7	8	9	10
Q bar	1.99	2.07	1.99	2.02	2.03	2.10	1.97	2.06	2.00	2.00
U bar	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
T	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
df	705.55	2183.9	863.25	1403.4	902.49	414.19	1293.2	1319.5	1274.9	1220.7
fmi	0.27	0.15	0.24	0.19	0.23	0.35	0.20	0.19	0.20	0.20
r	0.36	0.18	0.31	0.23	0.30	0.52	0.24	0.24	0.24	0.25
eff	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
t-val	1.96	1.96	1.96	1.96	1.96	1.97	1.96	1.96	1.96	1.96
95% CI	1.81	1.90	1.81	1.85	1.85	1.90	1.78	1.89	1.83	1.82
95% CI	2.18	2.24	2.18	2.20	2.21	2.29	2.15	2.24	2.17	2.17
width	0.37	0.35	0.37	0.36	0.36	0.39	0.37	0.36	0.34	0.36

Q bar –estimated mean for MI (Rubin's rules), U bar –estimated variance for MI (Rubin's rules), B – between-imputation variance, T – total variance, df – degrees of freedom, fmi – fraction of missing information, r – relative increase in variance, eff – efficiency rate, 95% CI – 95% confidence intervals, width – confidence interval width

MCAR HR Low	Variables									
MI	1	2	3	4	5	6	7	8	9	10
Q bar	2.09	2.03	2.03	1.98	1.99	2.05	1.91	1.94	1.92	2.01
U bar	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
T	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
df	908.44	642.44	1854.2	731.34	1163.2	1234.5	1090.5	817.52	626.73	837.13
fmi	0.23	0.28	0.16	0.26	0.21	0.20	0.21	0.25	0.28	0.24
r	0.30	0.38	0.19	0.35	0.26	0.25	0.27	0.32	0.39	0.32
eff	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
t-val	1.96	1.96	1.96	1.96	1.96	1.96	1.96	1.96	1.96	1.96
95% CI	1.91	1.84	1.85	1.79	1.82	1.87	1.73	1.76	1.73	1.82
95% CI	2.27	2.21	2.20	2.16	2.16	2.22	2.08	2.13	2.10	2.20
width	0.36	0.37	0.35	0.37	0.34	0.36	0.35	0.37	0.37	0.38

Q bar –estimated mean for MI (Rubin's rules), U bar –estimated variance for MI (Rubin's rules), B – between-imputation variance, T – total variance, df – degrees of freedom, fmi – fraction of missing information, r – relative increase in variance, eff – efficiency rate, 95% CI – 95% confidence intervals, width – confidence interval width

L.6 MCAR HNR high and low correlation structure

MCAR HNR High	Variables									
MI	1	2	3	4	5	6	7	8	9	10
Q bar	2.22	2.02	2.10	2.02	2.02	2.02	2.02	2.02	1.92	1.92
U bar	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
B	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
T	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
df	276.55	356.79	393.62	Large	Large	Large	Large	Large	290.39	339.02
fmi	0.43	0.37	0.36	0.00	0.00	0.00	0.00	0.00	0.41	0.38
r	0.73	0.59	0.55	0.00	0.00	0.00	0.00	0.00	0.70	0.61
eff	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
t-val	1.97	1.97	1.97	1.96	1.96	1.96	1.96	1.96	1.97	1.97
95% CI	2.00	1.83	1.89	1.86	1.86	1.86	1.86	1.86	1.72	1.71
95% CI	2.44	2.22	2.32	2.18	2.18	2.18	2.18	2.18	2.12	2.12
width	0.43	0.39	0.42	0.32	0.32	0.32	0.32	0.32	0.40	0.41

Q bar –estimated mean for MI (Rubin's rules), U bar –estimated variance for MI (Rubin's rules), B – between-imputation variance, T – total variance, df – degrees of freedom, fmi – fraction of missing information, r – relative increase in variance, eff – efficiency rate, 95% CI – 95% confidence intervals, width – confidence interval width, Large – refers to a very large value which will not be displayed

MCAR HNR Low	Variables									
MI	1	2	3	4	5	6	7	8	9	10
Q bar	1.93	2.12	2.03	2.02	2.02	2.02	2.02	2.02	2.02	2.00
U bar	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
T	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
df	426.77	299.10	437.17	Large	Large	Large	Large	Large	378.91	292.52
fmi	0.34	0.41	0.34	0.00	0.00	0.00	0.00	0.00	0.36	0.41
r	0.51	0.68	0.50	0.00	0.00	0.00	0.00	0.00	0.56	0.69
eff	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
t-val	1.97	1.97	1.97	1.96	1.96	1.96	1.96	1.96	1.97	1.97
95% CI	1.73	1.91	1.84	1.86	1.86	1.86	1.86	1.86	1.83	1.79
95% CI	2.13	2.34	2.22	2.18	2.18	2.18	2.18	2.18	2.21	2.22
width	0.40	0.43	0.39	0.32	0.32	0.32	0.32	0.32	0.39	0.43

Q bar –estimated mean for MI (Rubin's rules), U bar –estimated variance for MI (Rubin's rules), B – between-imputation variance, T – total variance, df – degrees of freedom, fmi – fraction of missing information, r – relative increase in variance, eff – efficiency rate, 95% CI – 95% confidence intervals, width – confidence interval width, Large – refers to a very large value which will not be displayed

L.7 MCAR LR high and low correlation structure

MCAR LR High	Variables									
MI	1	2	3	4	5	6	7	8	9	10
Q bar	2.04	2.07	2.00	2.04	2.04	1.99	2.04	2.05	2.00	2.00
U bar	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
T	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
df	1994	2259.7	2920.5	7789.8	10839.9	11584.	9308.3	13102.6	11907.5	10327.4
fmi	0.16	0.15	0.13	0.08	0.07	0.07	0.07	0.06	0.06	0.07
r	0.19	0.17	0.15	0.09	0.07	0.07	0.08	0.07	0.07	0.07
eff	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
t-val	1.96	1.96	1.96	1.96	1.96	1.96	1.96	1.96	1.96	1.96
95% CI	1.87	1.90	1.83	1.88	1.87	1.83	1.87	1.88	1.84	1.84
95% CI	2.21	2.24	2.17	2.21	2.20	2.16	2.20	2.21	2.17	2.16
width	0.34	0.33	0.34	0.34	0.33	0.33	0.33	0.33	0.33	0.32

Q bar –estimated mean for MI (Rubin's rules), U bar –estimated variance for MI (Rubin's rules), B – between-imputation variance, T – total variance, df – degrees of freedom, fmi – fraction of missing information, r – relative increase in variance, eff – efficiency rate, 95% CI – 95% confidence intervals, width – confidence interval width, Large – refers to a very large value which will not be displayed

MCAR LR Low	Variables									
MI	1	2	3	4	5	6	7	8	9	10
Q bar	2.00	2.06	2.06	2.00	2.03	2.06	2.04	1.98	1.99	1.97
U bar	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
T	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
df	7423.7	18669.8	3576	10419.6	3500.1	5287.8	2397.1	7138.3	2165.7	10089
fmi	0.08	0.05	0.12	0.07	0.12	0.10	0.14	0.08	0.15	0.07
r	0.09	0.05	0.13	0.07	0.13	0.11	0.17	0.09	0.18	0.07
eff	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
t-val	1.96	1.96	1.96	1.96	1.96	1.96	1.96	1.96	1.96	1.96
95% CI	1.83	1.89	1.89	1.83	1.86	1.90	1.87	1.81	1.81	1.81
95% CI	2.16	2.22	2.23	2.17	2.20	2.23	2.21	2.15	2.16	2.14
width	0.33	0.33	0.34	0.33	0.34	0.33	0.34	0.33	0.35	0.33

Q bar –estimated mean for MI (Rubin's rules), U bar –estimated variance for MI (Rubin's rules), B – between-imputation variance, T – total variance, df – degrees of freedom, fmi – fraction of missing information, r – relative increase in variance, eff – efficiency rate, 95% CI – 95% confidence intervals, width – confidence interval width, Large – refers to a very large value which will not be displayed

L.8 MCAR LNR high and low correlation structure

MCAR LNR High	Variables									
MI	1	2	3	4	5	6	7	8	9	10
Q bar	2.04	2.04	2.00	2.02	2.02	2.02	2.02	2.02	2.00	2.01
U bar	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
T	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
df	1973.2	1836	1798.7	Large	Large	Large	Large	Large	2959.9	1459.8
fmi	0.16	0.16	0.17	0.00	0.00	0.00	0.00	0.00	0.13	0.18
r	0.19	0.20	0.20	0.00	0.00	0.00	0.00	0.00	0.15	0.22
eff	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
t-val	1.96	1.96	1.96	1.96	1.96	1.96	1.96	1.96	1.96	1.96
95% CI	1.87	1.87	1.82	1.86	1.86	1.86	1.86	1.86	1.83	1.84
95% CI	2.22	2.21	2.18	2.18	2.18	2.18	2.18	2.18	2.16	2.19
width	0.35	0.34	0.36	0.32	0.32	0.32	0.32	0.32	0.34	0.35

Q bar –estimated mean for MI (Rubin's rules), U bar –estimated variance for MI (Rubin's rules), B – between-imputation variance, T – total variance, df – degrees of freedom, fmi – fraction of missing information, r – relative increase in variance, eff – efficiency rate, 95% CI – 95% confidence intervals, width – confidence interval width, Large – refers to a very large value which will not be displayed

MCAR LNR Low	Variables									
MI	1	2	3	4	5	6	7	8	9	10
Q bar	2.06	2.03	1.97	2.02	2.02	2.02	2.02	2.02	2.00	1.98
U bar	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
T	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
df	2308.3	1787.6	1444.0	Large	Large	Large	Large	Large	2341.9	1768.7
fmi	0.15	0.17	0.19	0.00	0.00	0.00	0.00	0.00	0.15	0.17
r	0.17	0.20	0.23	0.00	0.00	0.00	0.00	0.00	0.17	0.20
eff	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
t-val	1.96	1.96	1.96	1.96	1.96	1.96	1.96	1.96	1.96	1.96
95% CI	1.88	1.85	1.79	1.86	1.86	1.86	1.86	1.86	1.83	1.81
95% CI	2.23	2.21	2.15	2.18	2.18	2.18	2.18	2.18	2.17	2.16
width	0.35	0.35	0.36	0.32	0.32	0.32	0.32	0.32	0.34	0.36

Q bar –estimated mean for MI (Rubin's rules), U bar –estimated variance for MI (Rubin's rules), B – between-imputation variance, T – total variance, df – degrees of freedom, fmi – fraction of missing information, r – relative increase in variance, eff – efficiency rate, 95% CI – 95% confidence intervals, width – confidence interval width, Large – refers to a very large value which will not be displayed

Appendix M: Rubin's rules results for real data

M.1 RIMCA

RIMCA	Variables						
MI	1	2	3	4	5	6	7
Q bar	1.4301	1.9839	1.2926	1.2675	1.2577	1.1611	1.4746
U bar	0.0003	0.0006	0.0003	0.0002	0.0002	0.0001	0.0002
B	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
T	0.0004	0.0006	0.0003	0.0002	0.0002	0.0001	0.0002
df	52548.90	15486.17	75691.62	63364.36	111631.60	160231.13	80662.69
fmi	0.0306	0.0564	0.0255	0.0278	0.0210	0.0175	0.0247
r	0.0315	0.0596	0.0261	0.0286	0.0214	0.0178	0.0253
eff	0.9997	0.9994	0.9997	0.9997	0.9998	0.9998	0.9998
t-val	1.9600	1.9601	1.9600	1.9600	1.9600	1.9600	1.9600
95% CI	1.3935	1.9349	1.2607	1.2425	1.2330	1.1404	1.4463
95% CI	1.4668	2.0328	1.3244	1.2926	1.2824	1.1818	1.5028
width	0.0734	0.0980	0.0637	0.0502	0.0494	0.0414	0.0565
RIMCA	**Variables**						
MI	8	9	10	11	12	13	14
Q bar	1.4082	1.9390	1.4033	1.4506	2.0157	2.1937	1.7153
U bar	0.0002	0.0003	0.0002	0.0002	0.0003	0.0004	0.0003
B	0.0000	0.0000	0.0000	0.0000	0.0001	0.0000	0.0000
T	0.0002	0.0004	0.0002	0.0003	0.0003	0.0004	0.0003
df	22475.93	5096.20	1504.19	1232.12	1064.91	71154.90	59993.78
fmi	0.0468	0.0984	0.1816	0.2007	0.2160	0.0263	0.0286
r	0.0490	0.1087	0.2202	0.2491	0.2731	0.0269	0.0294
eff	0.9995	0.9990	0.9982	0.9980	0.9978	0.9997	0.9997
t-val	1.9601	1.9604	1.9615	1.9619	1.9622	1.9600	1.9600
95% CI	1.3801	1.9013	1.3730	1.4196	1.9792	2.1523	1.6837
95% CI	1.4364	1.9767	1.4336	1.4817	2.0522	2.2350	1.7469
width	0.0562	0.0754	0.0606	0.0622	0.0731	0.0827	0.0632

Q bar –estimated mean for MI (Rubin's rules), U bar –estimated variance for MI (Rubin's rules), B – between-imputation variance, T – total variance, df – degrees of freedom, fmi – fraction of missing information, r – relative increase in variance, eff – efficiency rate, 95% CI – 95% confidence intervals, width – confidence interval width

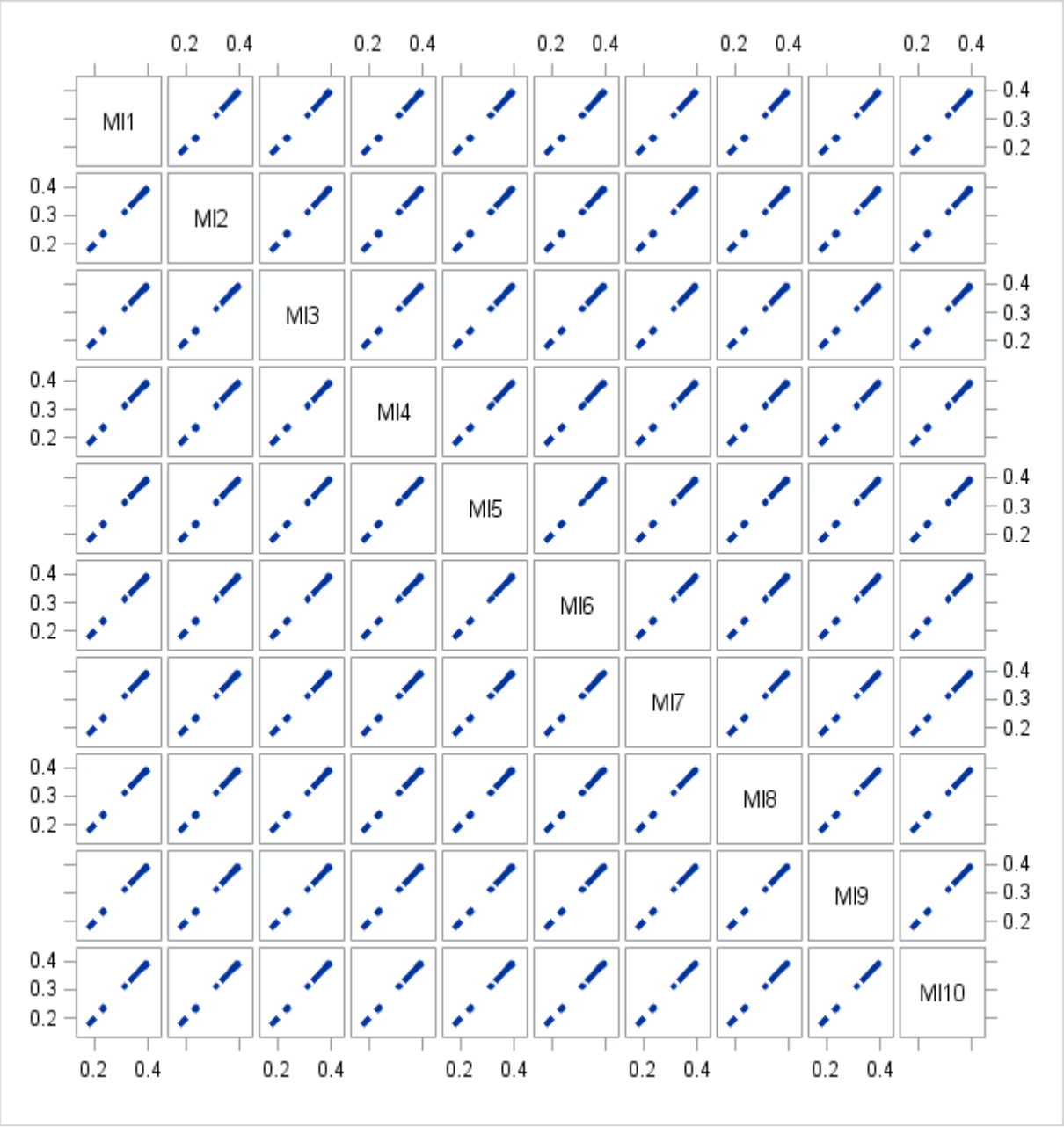
M.2 IMCA

IMCA	Variables						
MI	1	2	3	4	5	6	7
Q bar	1.4296	1.9864	1.2928	1.2676	1.2573	1.1609	1.4743
U bar	0.0003	0.0006	0.0003	0.0002	0.0002	0.0001	0.0002
B	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
T	0.0004	0.0006	0.0003	0.0002	0.0002	0.0001	0.0002
df	33615.62	11528.99	39776.97	82719.17	146527.63	165406.02	30939.76
fmi	0.0382	0.0654	0.0351	0.0244	0.0183	0.0172	0.0399
r	0.0397	0.0697	0.0364	0.0249	0.0186	0.0175	0.0414
eff	0.9996	0.9993	0.9996	0.9998	0.9998	0.9998	0.9996
t-val	1.9600	1.9602	1.9600	1.9600	1.9600	1.9600	1.9600
95% CI	1.3929	1.9372	1.2607	1.2426	1.2326	1.1402	1.4458
95% CI	1.4664	2.0356	1.3248	1.2927	1.2819	1.1816	1.5027
width	0.0736	0.0984	0.0641	0.0501	0.0493	0.0414	0.0569
IMCA	**Variables**						
MI	8	9	10	11	12	13	14
Q bar	1.4087	1.9392	1.3993	1.4501	2.0150	2.1943	1.7150
U bar	0.0002	0.0003	0.0002	0.0002	0.0003	0.0004	0.0003
B	0.0000	0.0000	0.0001	0.0000	0.0001	0.0000	0.0000
T	0.0002	0.0004	0.0002	0.0002	0.0004	0.0004	0.0003
df	12941.55	3857.23	1026.67	1477.10	651.10	63963.11	50638.61
fmi	0.0617	0.1132	0.2200	0.1832	0.2766	0.0277	0.0311
r	0.0656	0.1270	0.2795	0.2227	0.3780	0.0285	0.0321
eff	0.9994	0.9989	0.9978	0.9982	0.9972	0.9997	0.9997
t-val	1.9601	1.9606	1.9623	1.9616	1.9636	1.9600	1.9600
95% CI	1.3803	1.9012	1.3683	1.4194	1.9769	2.1528	1.6834
95% CI	1.4370	1.9772	1.4302	1.4809	2.0530	2.2357	1.7466
width	0.0567	0.0761	0.0620	0.0615	0.0761	0.0828	0.0632

Q bar –estimated mean for MI (Rubin's rules), U bar –estimated variance for MI (Rubin's rules), B – between-imputation variance, T – total variance, df – degrees of freedom, fmi – fraction of missing information, r – relative increase in variance, eff – efficiency rate, 95% CI – 95% confidence intervals, width – confidence interval width

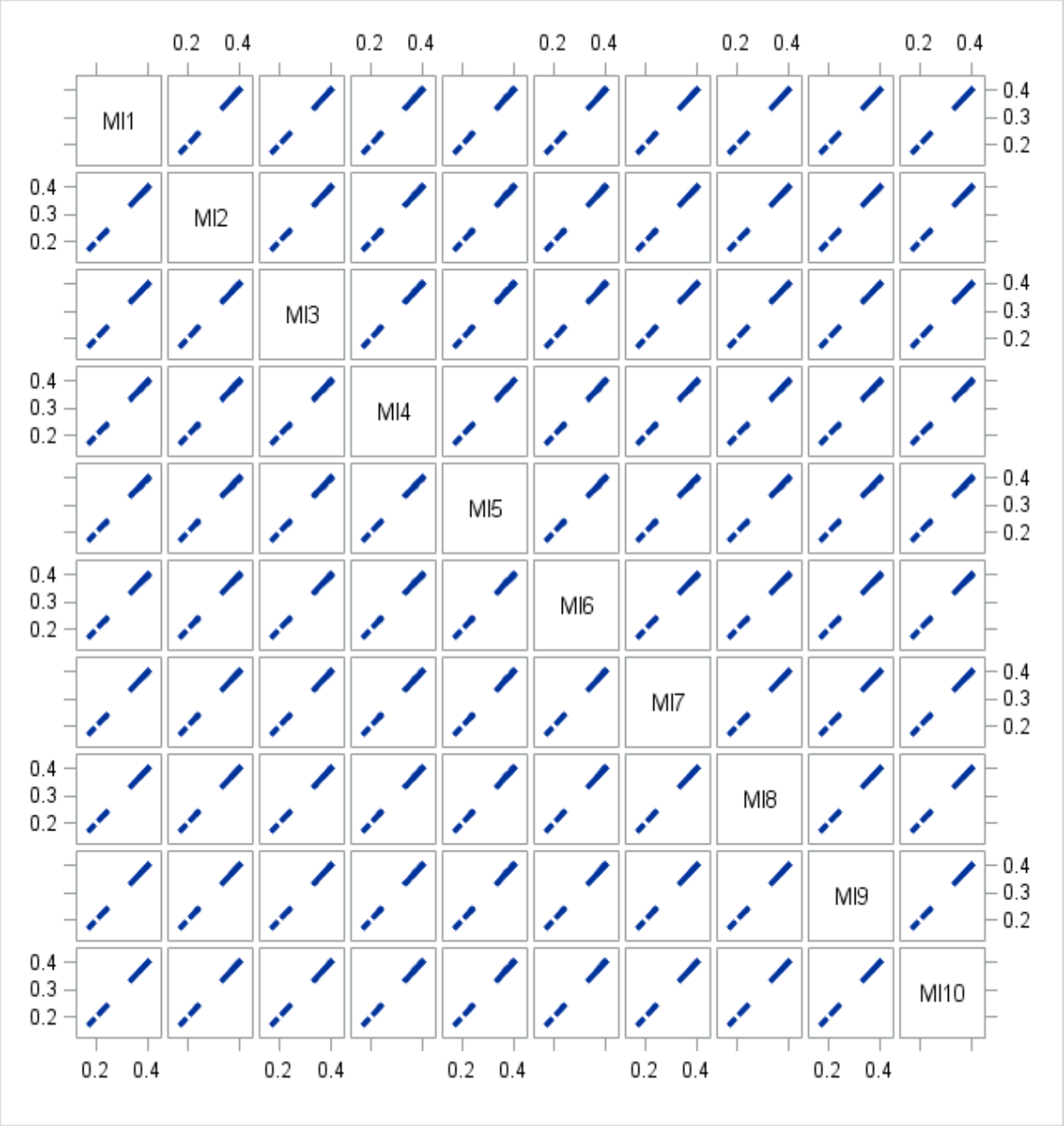
Appendix N: Scatterplot matrices

N.1 MAR HR with high correlation structure



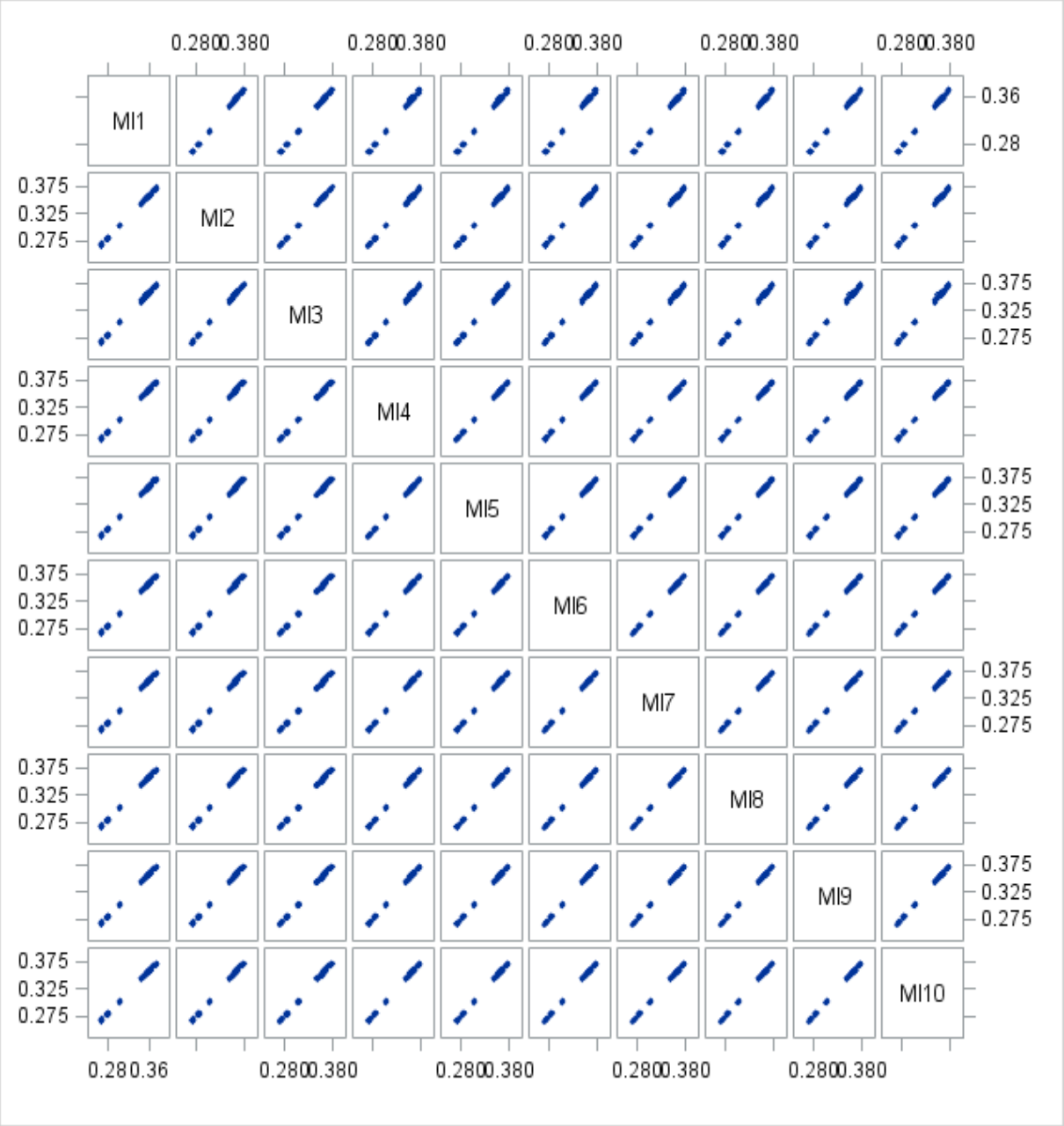
Generated by the SAS System ('Local', X64_7PRO) on 10 July 2013 at 2:42:41 PM

N.2 MAR HNR with high correlation structure



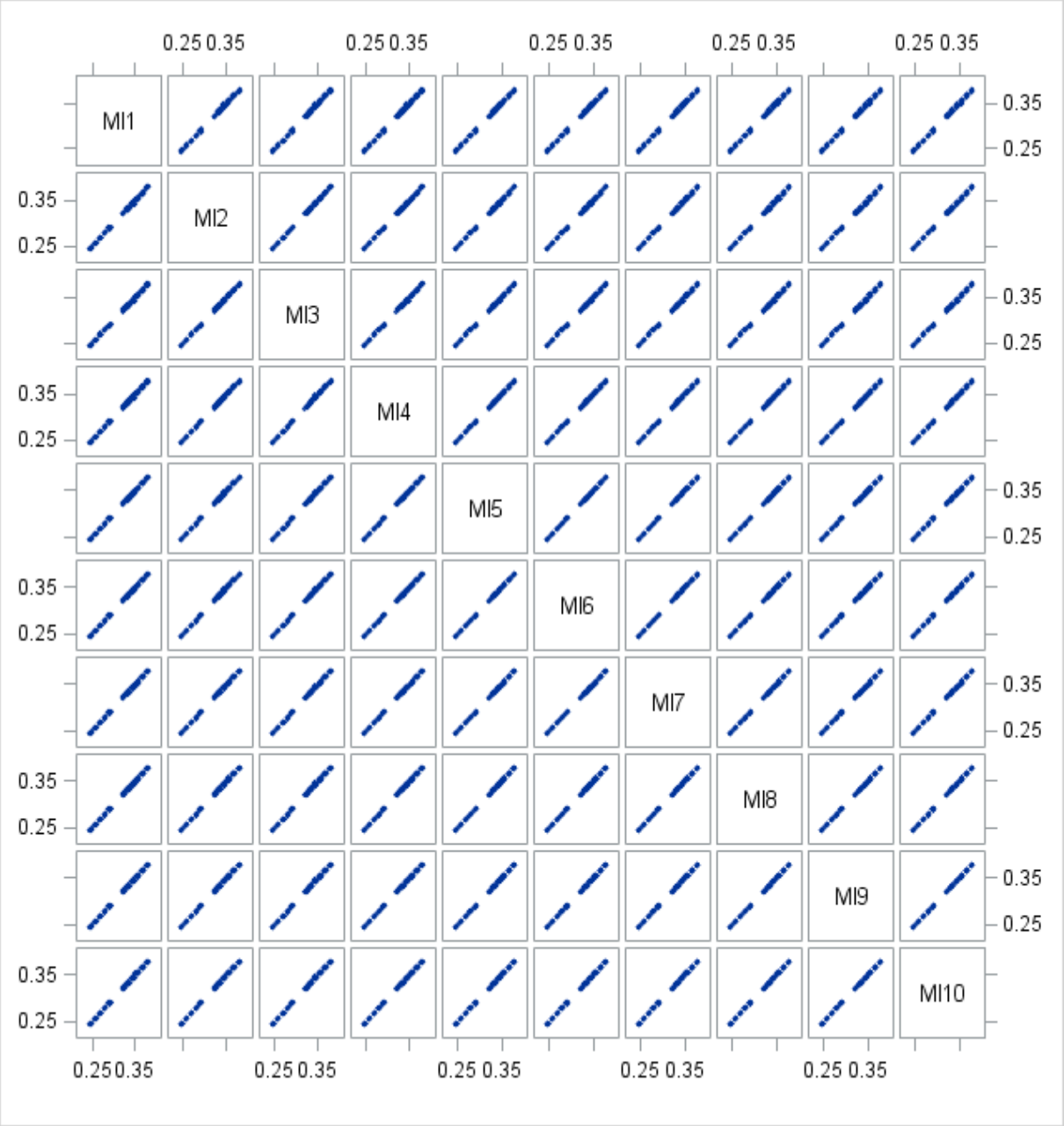
Generated by the SAS System ('Local', X64_7PRO) on 10 July 2013 at 2:44:26 PM

N.3 MAR LR with high correlation structure



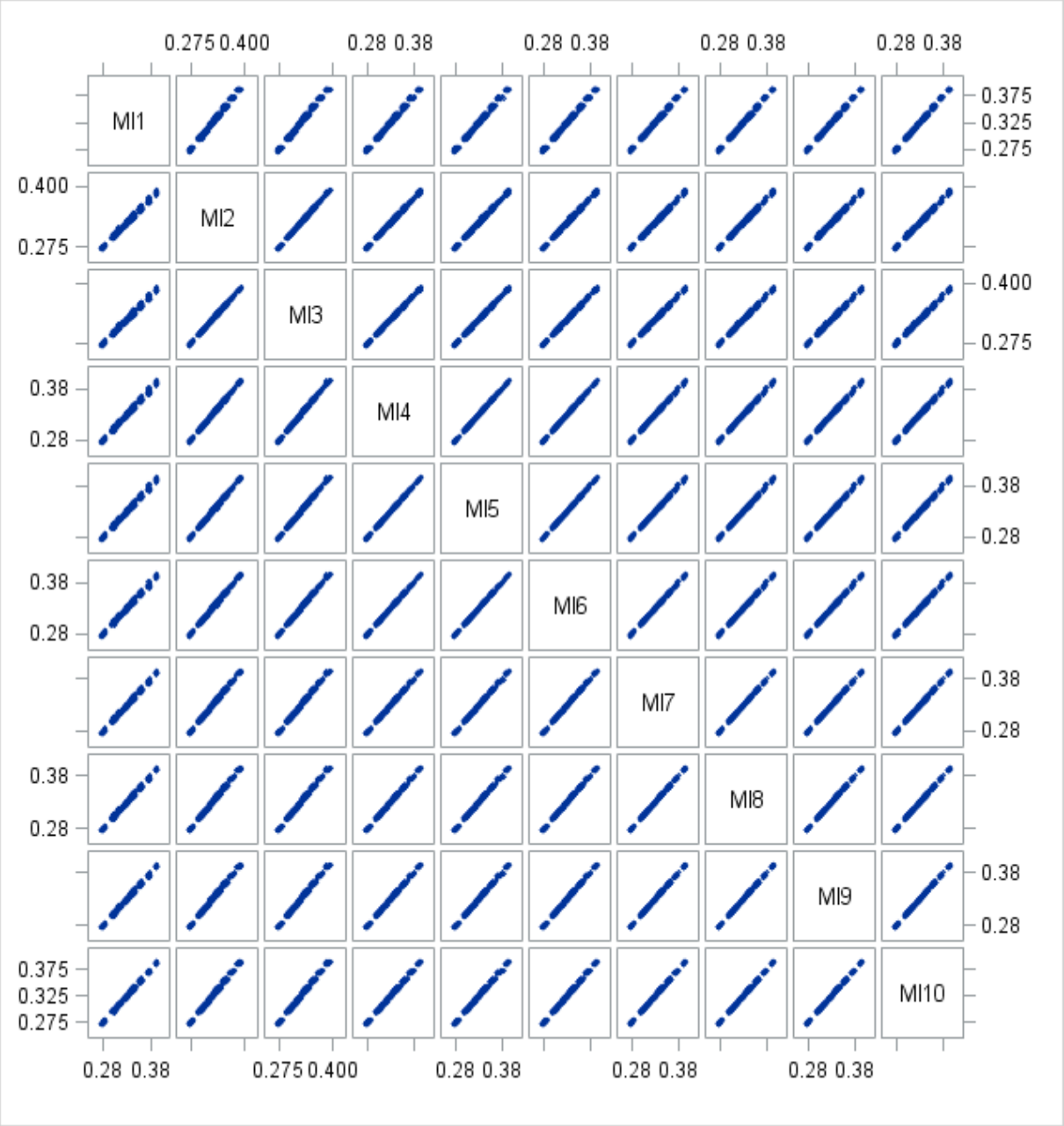
Generated by the SAS System ('Local', X64_7PRO) on 10 July 2013 at 2:39:15 PM

N.4 MAR LNR with high correlation structure



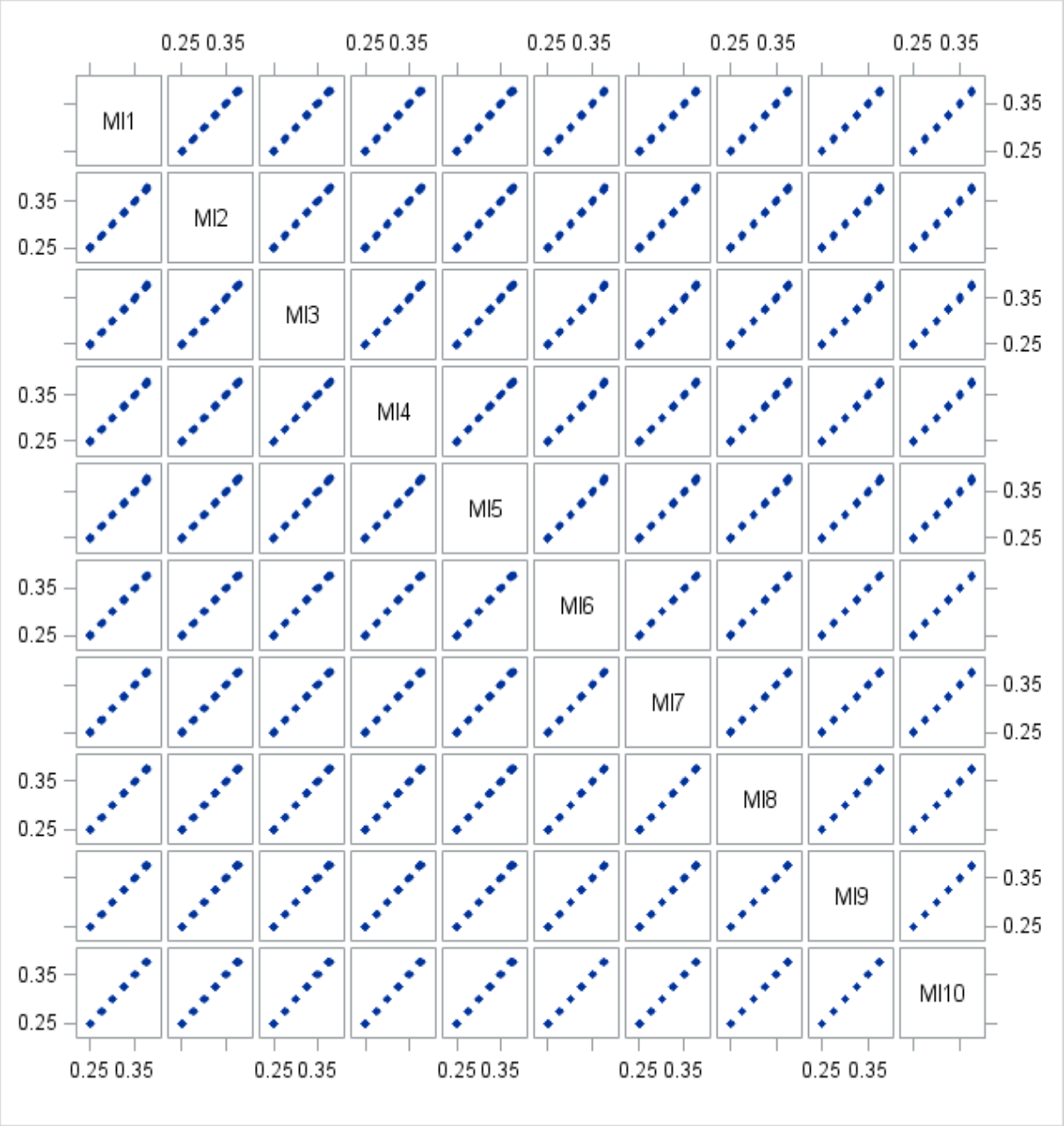
Generated by the SAS System ('Local', X64_7PRO) on 10 July 2013 at 2:40:55 PM

N.5 MCAR HR with high correlation structure



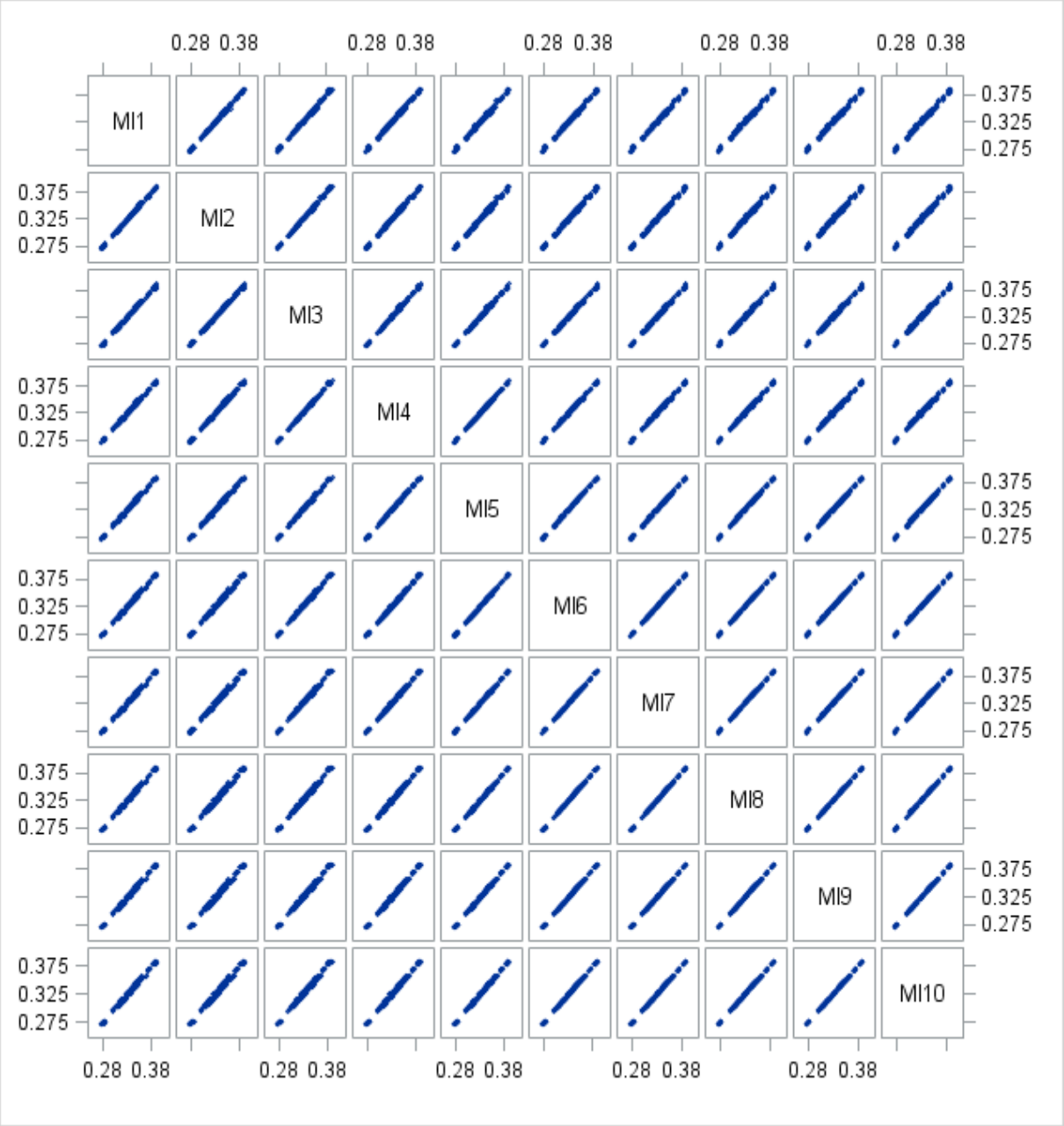
Generated by the SAS System ('Local', X64_7PRO) on 10 July 2013 at 1:05:12 PM

N.6 MCAR HNR with high correlation structure



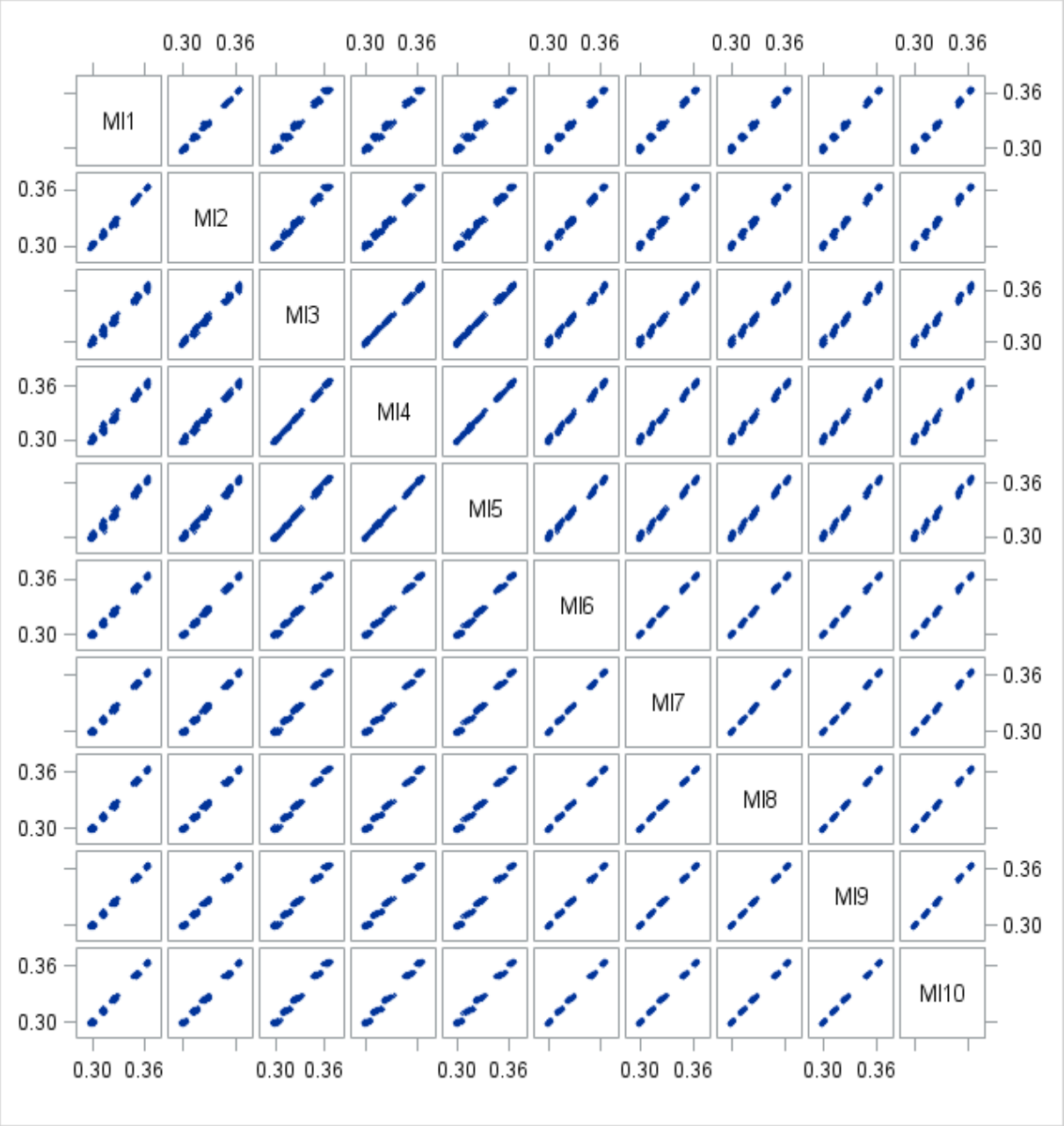
Generated by the SAS System ('Local', X64_7PRO) on 10 July 2013 at 1:07:56 PM

N.7 MCAR LR with high correlation structure



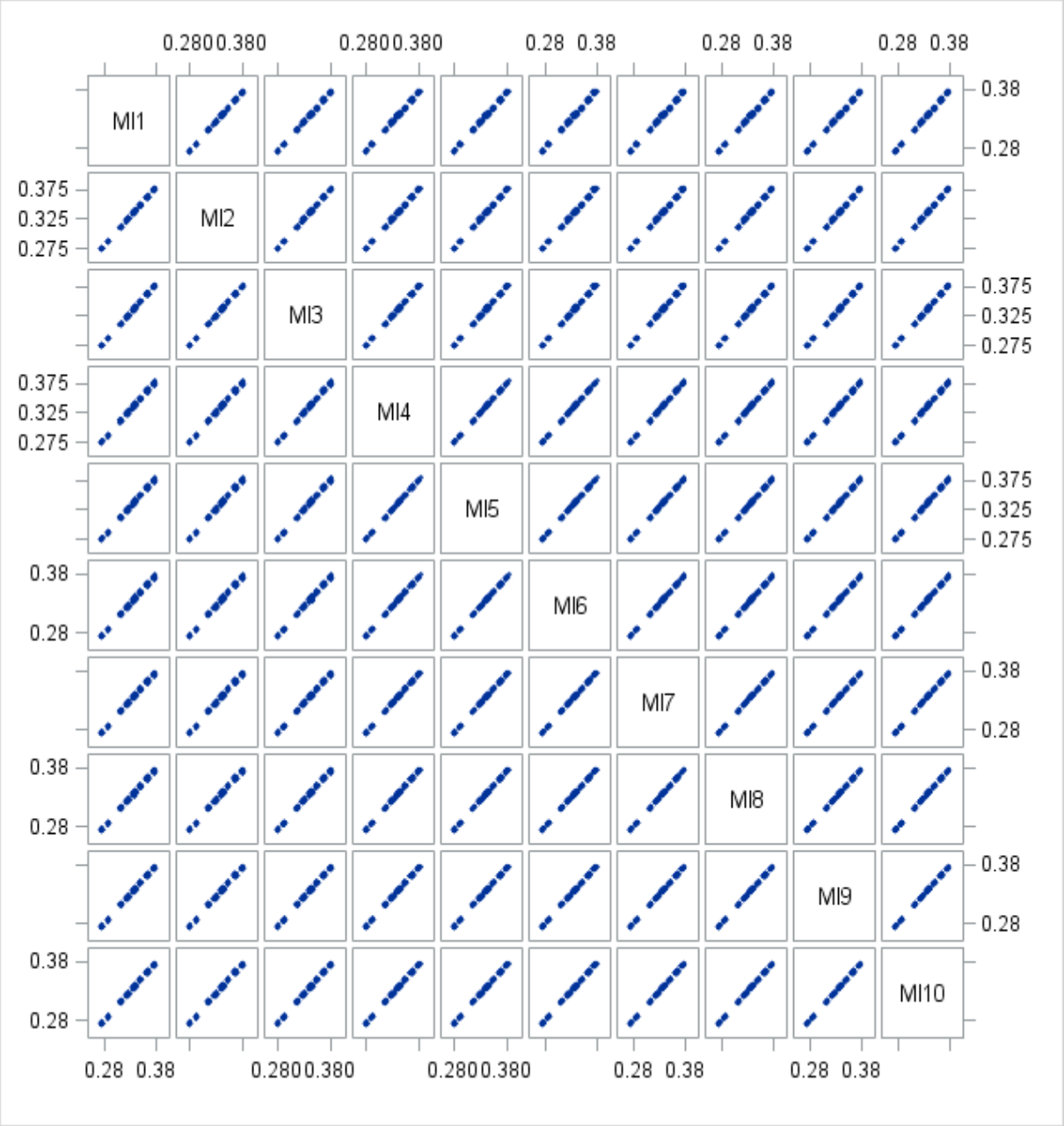
Generated by the SAS System ('Local', X64_7PRO) on 10 July 2013 at 12:37:56 PM

N.8 MCAR LNR with high correlation structure



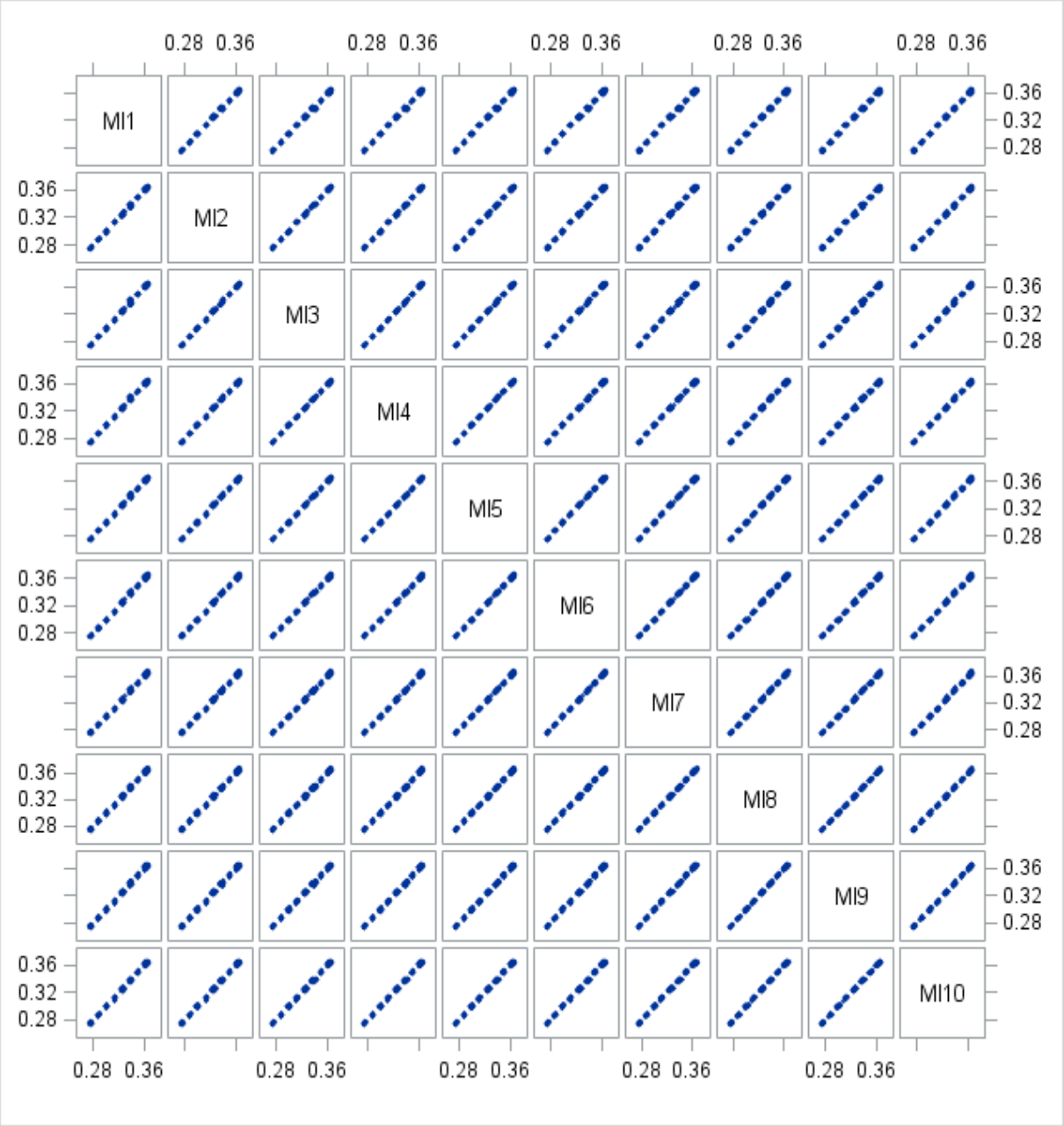
Generated by the SAS System ('Local', X64_7PRO) on 10 July 2013 at 1:01:05 PM

N.9 MAR HR with low correlation structure



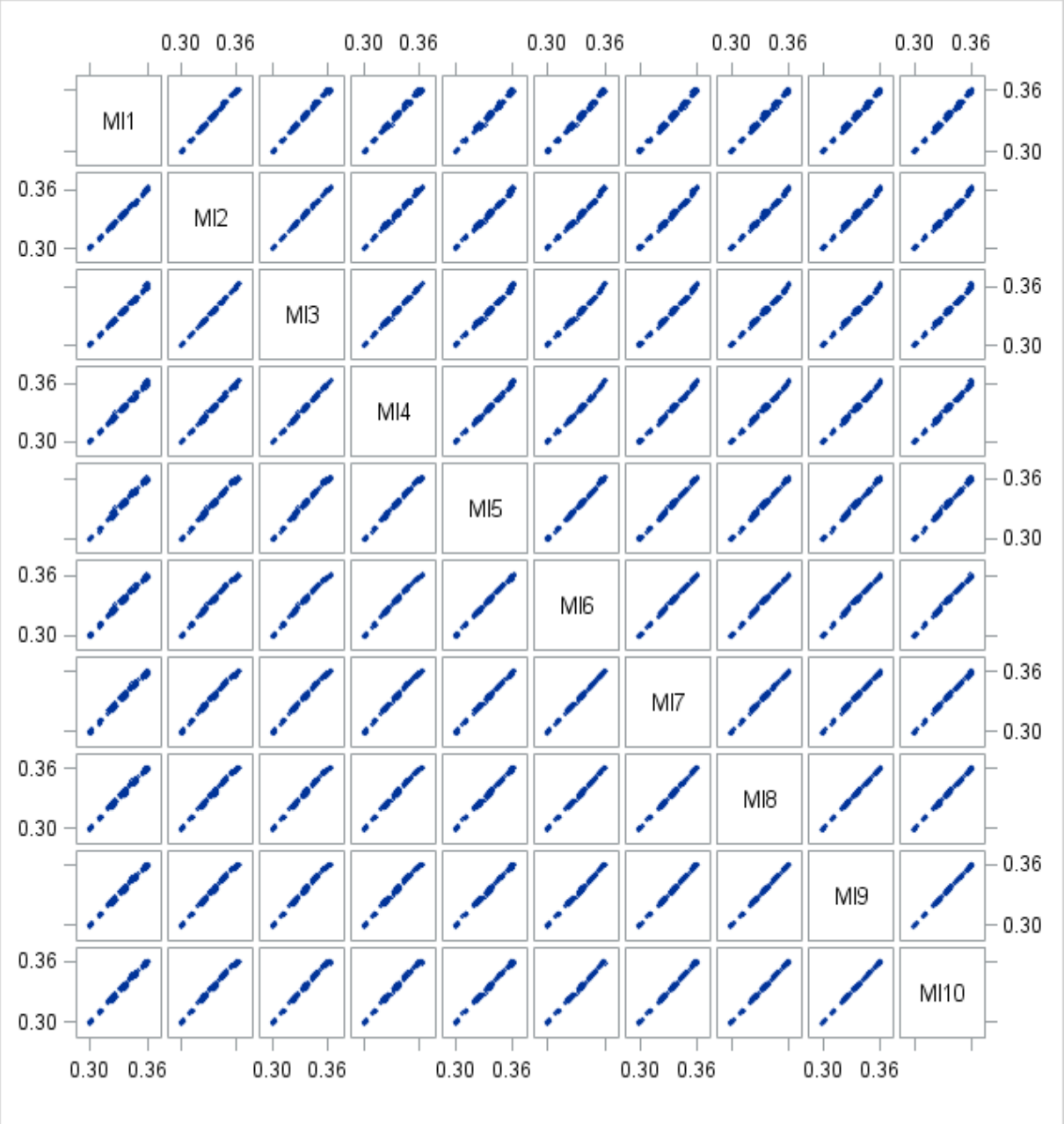
Generated by the SAS System ('Local', X64_7PRO) on 10 July 2013 at 3:12:06 PM

N.10 MAR HNR with low correlation structure



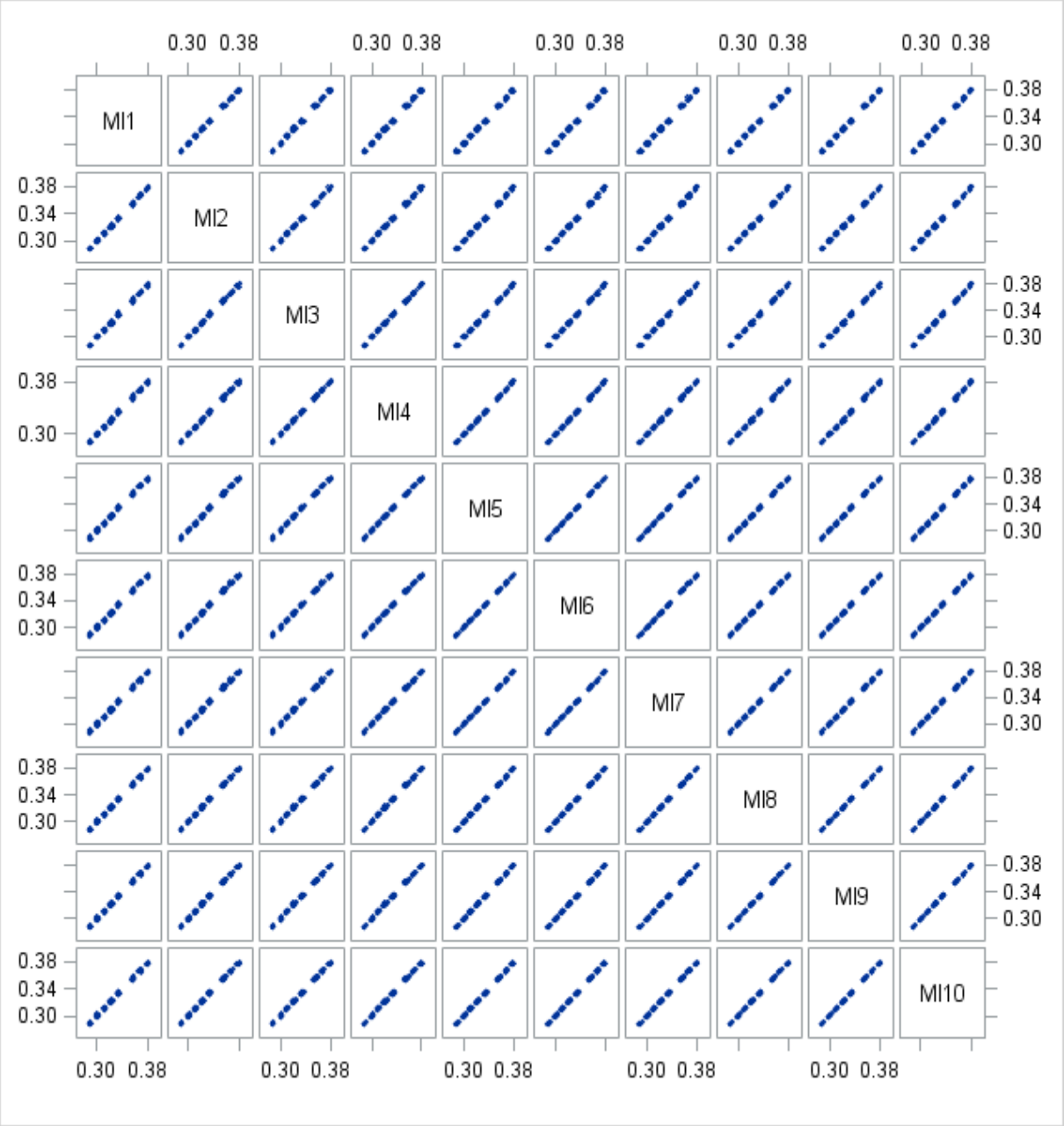
Generated by the SAS System ('Local', X64_7PRO) on 10 July 2013 at 3:13:48 PM

N.11 MAR LR with low correlation structure



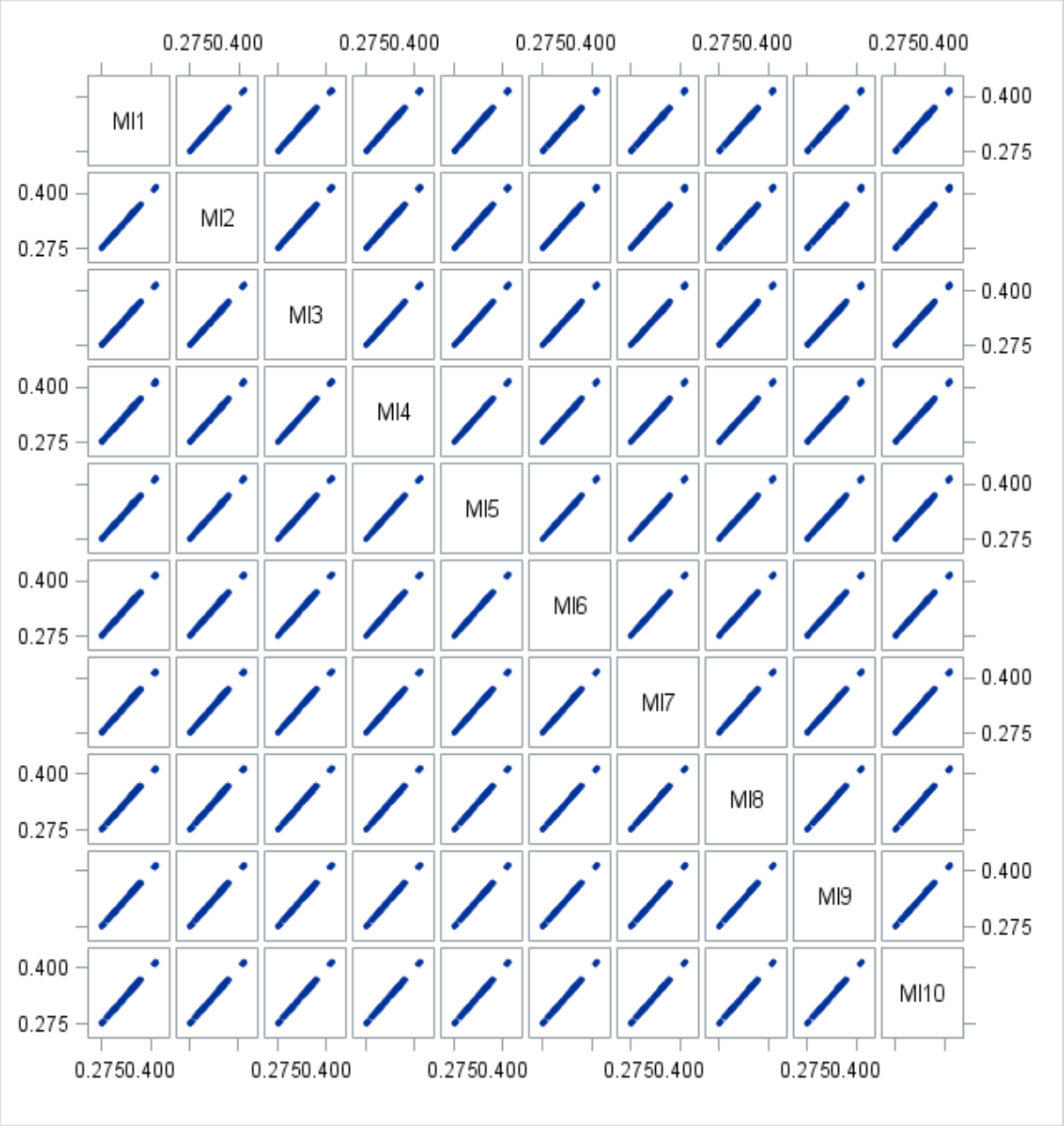
Generated by the SAS System ('Local', X64_7PRO) on 10 July 2013 at 3:08:16 PM

N.12 MAR LNR with low correlation structure



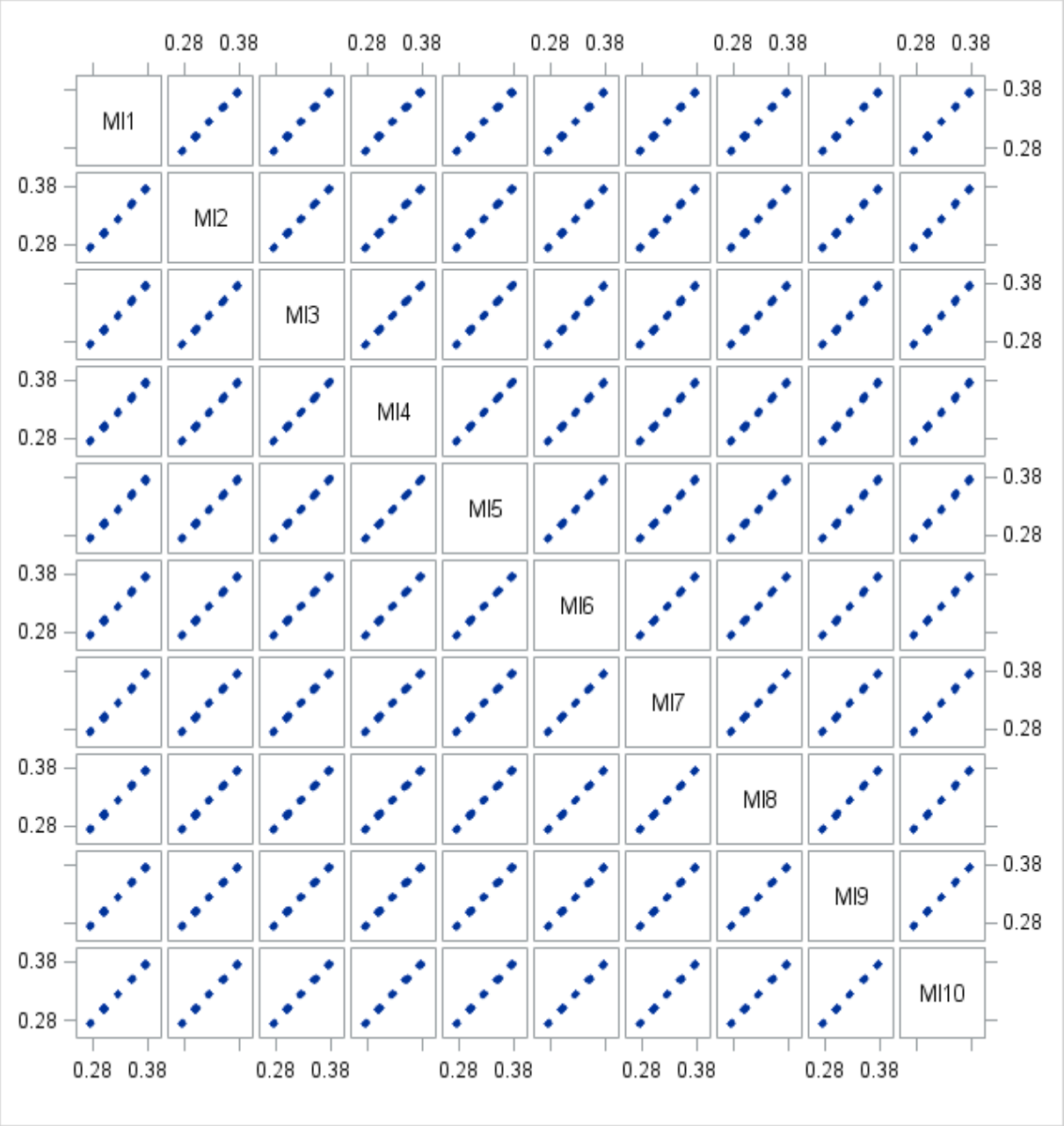
Generated by the SAS System ('Local', X64_7PRO) on 10 July 2013 at 3:10:22 PM

N.13 MCAR HR with low correlation structure



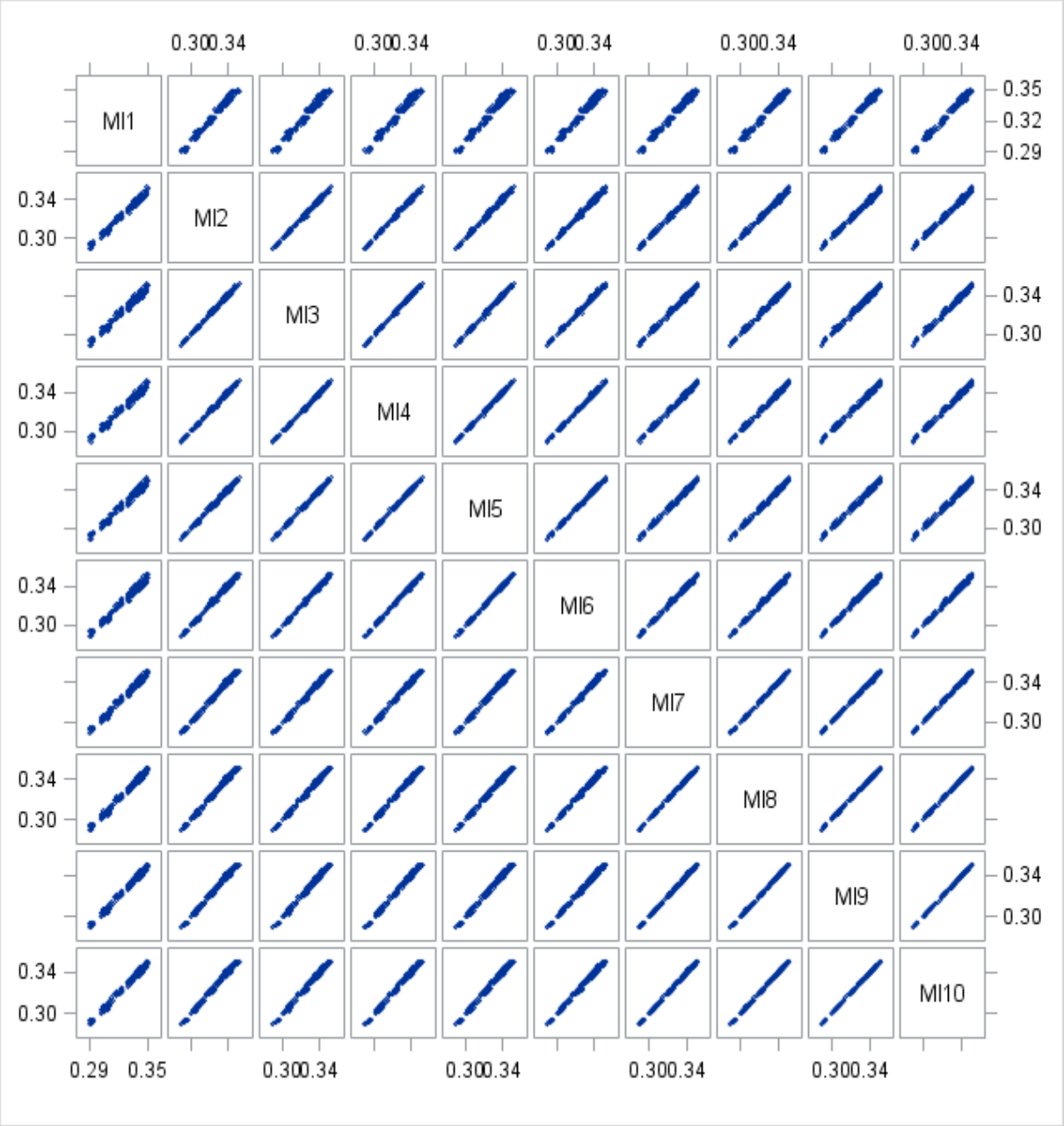
Generated by the SAS System ('Local', X64_7PRO) on 10 July 2013 at 1:35:12 PM

N.14 MCAR HNR with low correlation structure



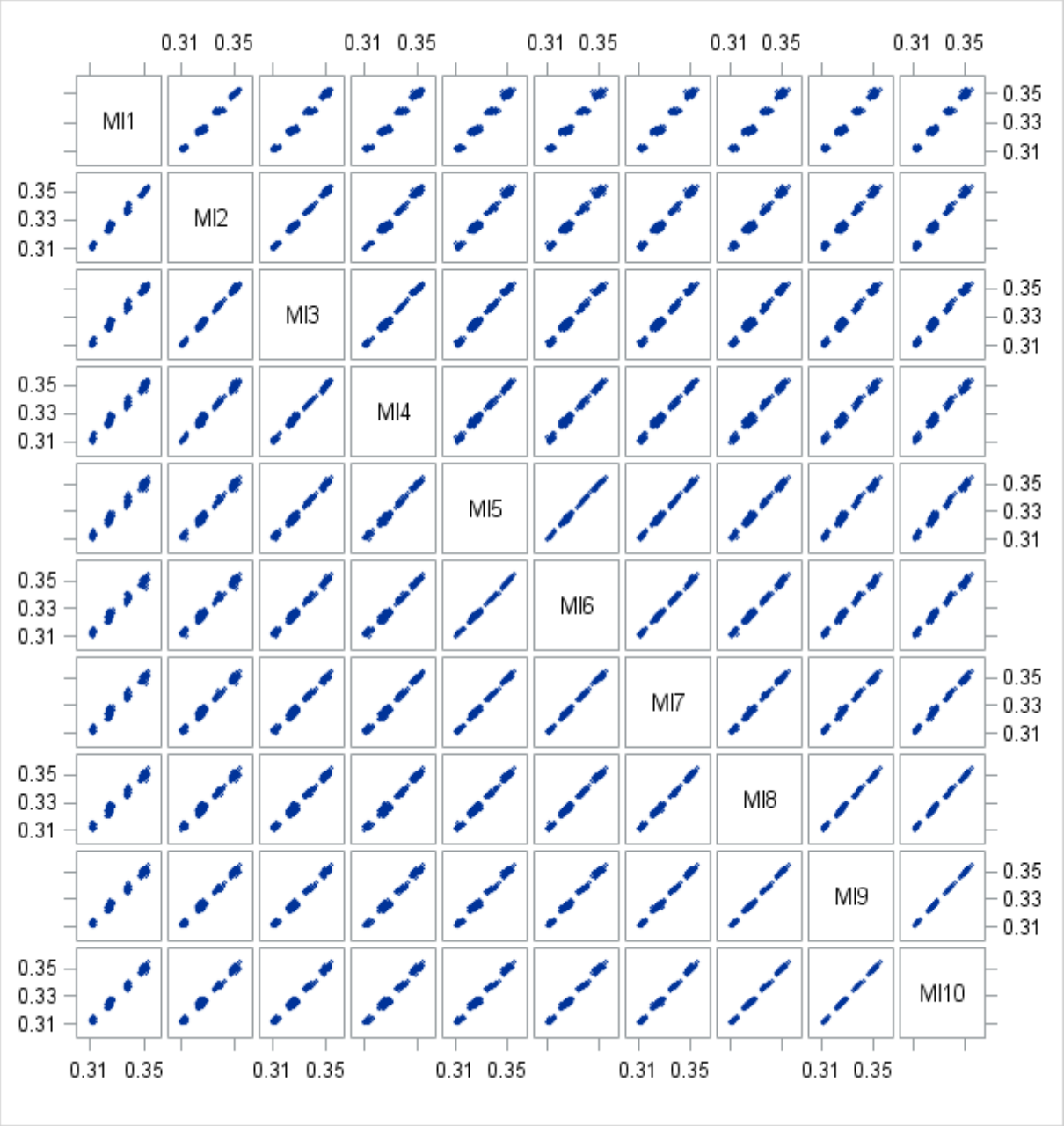
Generated by the SAS System ('Local', X64_7PRO) on 10 July 2013 at 1:37:02 PM

N.15 MCAR LR with low correlation structure



Generated by the SAS System ('Local', X64_7PRO) on 10 July 2013 at 1:31:32 PM

N.16 MCAR LNR with low correlation structure



Generated by the SAS System ('Local', X64_7PRO) on 10 July 2013 at 1:33:18 PM

Summary

Key terms: incomplete ordinal categorical data, missingness mechanisms, multiple imputation, multiple correspondence analysis, principal component analysis, regularised iterative multiple correspondence analysis.

Non-responses in survey data are a prevalent problem. Various techniques for the handling of missing data have been studied and published. The application of a regularised iterative multiple correspondence analysis (RIMCA) algorithm in single imputation (SI) has been suggested for the handling of missing data in survey analysis.

Multiple correspondence analysis (MCA) as an imputation procedure is appropriate for survey data, since MCA is concerned with the relationships among the variables in the data. Therefore, missing data can be imputed by exploiting the relationship between observed and missing data.

The RIMCA algorithm expresses MCA as a weighted principal component analysis (PCA) of a data triplet (Z, M, D), which represents a weighted data matrix, a metric and a diagonal matrix containing row masses, respectively. Performing PCA on a triplet involves the generalised singular value decomposition of the weighted data matrix Z. Here, standard singular value decomposition (SVD) will not suffice, since constraints are imposed on the rows and columns because of the weighting.

The success of this algorithm lies in the fact that all eigenvalues are shrunk and the last components are omitted; thus a 'double shrinkage' occurs, which reduces variance and stabilises predictions. RIMCA seems to overcome overfitting and underfitting problems with regard to categorical missing data in surveys.

The idea of applying the RIMCA algorithm in MI was appealing, since advantages of MI occur over SI, such as an increase in the accuracy of

estimations and the attainment of valid inferences when combining multiple datasets.

The aim of this study was to establish the performance of RIMCA in MI. This was achieved by two objectives: to determine whether RIMCA in MI outperforms RIMCA in SI and to determine the accuracy of predictions made from RIMCA in MI as an imputation model.

Real and simulated data were used. A simulation protocol was followed creating data drawn from multivariate Normal distributions with both high and low correlation structures. Varying the percentages of missing values in the data and missingness mechanisms (missing completely at random (MCAR) and missing at random (MAR)), as is done by Josse *et al.* (2012), were created in the data.

The first objective was achieved by applying RIMCA in both SI and MI to real data and simulated data. The performance of RIMCA in SI and MI were compared with regard to the obtained mean estimates and confidence intervals. In the case of the real data, the estimates were compared to the mean estimates of the incomplete data, whereas for the simulated data the true mean values and confidence intervals could be compared to the estimates obtained from the imputation procedures.

The second objective was achieved by calculating the apparent error rates of predictions made by the RIMCA algorithm in SI and MI in simulated datasets. Along with the apparent error rates, approximate overall success rates were calculated in order to establish the accuracy of imputations made by the SI and MI.

The results of this study show that the confidence intervals provided by MI are wider in most of the cases, which confirmed the incorporation of additional variance. It was found that for some of the variables the SI procedures were statistically different from the true confidence intervals, which shows that SI was not suitable in these instances for imputation. Overall the mean estimates provided by MI were closer to the true values, with respect to the simulated and real data. A summary of the bias, mean square errors and coverage for the

imputation techniques over a thousand simulations were provided, which also confirmed that RIMCA in MI was a better model than RIMCA in SI in the contexts provided by this research.

Opsomming

Sleutelwoorde: ontbrekende ordinale kategoriese data, verlorenis meganismes, meervoudige imputasie, meervoudige ooreenkomsanalise, hoofkomponentanalise, regulariseerde iteratiewe meervoudige ooreenkomsanalise.

Die verskynsel van ontbrekende waardes in vraelyste is 'n algemene probleem. Verskeie tegnieke vir die hantering van ontbrekende waardes is gebestudeer en gepubliseer. Die toepassing van 'n regulariseerde iteratiewe meervoudige ooreenkomsanalise (RIMCA) algoritme in enkelvoudige imputasie is voorgestel vir die hantering van ontbrekende waardes in die konteks van vraelyste.

Meervoudige ooreenkomsanalise (MCA) as 'n imputasie prosedure is gepas vir vraelys data, aangesien MCA die verhoudings tussen veranderlikes in die data benut. Dus kan die ontbrekende waardes opgevul word deur imputasie wat bepaal word deur die verhoudings tussen die waargenome en ontbrekende data.

Die RIMCA algoritme omskryf MCA as 'n geweegde hoofkomponentanalise (PCA) wat die data as 'n drietal (Z, M, D) uitdruk. Die drietal stel die geweegde data, metries en ry massas, onderskeidelik voor. Die uitvoer van PCA op 'n drietal sluit die toepassing van 'n veralgemeende singuliere waarde-ontbinding van die geweegde data matriks Z in. Standaard singuliere waarde-ontbinding is nie voldoende in hierdie geval nie, aangesien beperkings op die rye en kolomme geplaas word as gevolg van die geweegde data matriks.

Die sukses van hierdie algoritme is die verkleining van die eiewaardes en die weglaat van die laaste komponente. Dus ontstaan daar 'n dubbelle krimping wat sodoende die variansie laat afneem en voorspellings stabiliseer. Dit wil voorkom asof RIMCA oormatige passings- en ondermatige passingsprobleme in die konteks van kategoriese ontbrekende waardes in vraelyste oorkom.

Die idee vir die toepassing van RIMCA in meervoudige imputasie was aanloklik, aangesien meervoudige imputasie voordele inhou bo enkelvoudige imputasie met ‘n toename in die akkuraatheid van beramings en die verkryging van geldige inferensie wanneer die meervoudige datastelle saamgevoeg word.

Die doel van hierdie studie was om die prestasie van RIMCA in meervoudige imputasie te evalueer. Daar was twee doelstellings, naamlik: om vas te stel of RIMCA in meervoudige imputasie beter vaar as RIMCA in enkelvoudige imputasie, asook om die akkuraatheid van voorspellings gemaak deur RIMCA in meervoudige imputasie vas te stel.

Werklike en gesimuleerde data is gebruik. ‘n Simulasie protokol is gevolg wat gebruik is deur Josse *et al.* (2012) waarin waardes van ‘n meerveranderlike Normaal verdeling met hoë en lae korrelasie struktuur geneem is. Ontbrekende waardes is in die volledige datastelle geplaas volgens verskillende vereistes aangaande die persentasie ontbrekende waardes in die data, sowel as die tipe verlorenis meganisme (algeheel stogasties verlore (MCAR) en stogasties verlore (MAR)).

Die eerste doelstelling is bereik deur die toepassing van RIMCA in beide enkelvoudige en meervoudige imputasie op ware data en gesimuleerde data. Die optrede van RIMCA in enkelvoudige imputasie is vergelyk met dié van meervoudige imputasie deur middel van gemiddelde beramings en vertrouensintervalle. In die geval van die werklike data is die imputasie beramings met die ontbrekende beramings vergelyk, terwyl die gesimuleerde data die navorser toegelaat het om die beraamde imputasie waardes met die ware gemiddelde waardes te kon vergelyk.

Die tweede doelstelling is bereik deur die berekening van die skynbare foutkoers van die voorspellings gemaak deur die RIMCA algoritme in enkelvoudige- en meervoudige imputasie. Die benaderde algehele sukseskoers is bereken om sodoende die akkuraatheid van die imputasies deur beide enkelvoudige- en meervoudige imputasie te bepaal, asook die skynbare foutkoerse.

Die resultate van hierdie studie het aangedui dat die vertrouensintervalle verkry van die meervoudige imputasie tegniek wyer was as die intervalle verkry van die enkelvoudige imputasie tegniek. Hierdie bevinding bevestig die addisionele onsekerheid wat deur meervoudige imputasie bygevoeg word. Sommige van die veranderlikes het statisties verskil van die ware vertrouensintervalle na die toepassing van enkelvoudige imputasie, en daarom was enkelvoudige imputasie nie in hierdie gevalle geskik nie. In die algemeen was die gemiddelde beramings van meervoudige imputasie nader aan die ware gemiddelde waardes in beide die ware- en gesimuleerde data. ‘n Opsomming van die sydigheid, gemiddelde kwadratiese fout en die dekking van die imputasie tegnieke oor ‘n duisend simulasies het tesame met die doelstellingsresultate bevestig dat RIMCA in meervoudige analise ‘n beter en meer gepaste model is as RIMCA in enkelvoudige imputasie.