Multiple Imputation and Access to Likelihood Based Tools in Missing Data Problems

Firouzeh Noghrehchi

School of Mathematics and Statistics
The University of New South Wales

December, 2018

Submitted in total fulfillment of the requirements
of the degree of Doctor of Philosophy

Originality statement

'I hereby declare that this submission is my own work and to the best of my knowledge it contains no materials previously published or written by another person, or substantial proportions of material which have been accepted for the award of any other degree or diploma at UNSW or any other educational institution, except where due acknowledgement is made in the thesis. Any contribution made to the research by others, with whom I have worked at UNSW or elsewhere, is explicitly acknowledged in the thesis. I also declare that the intellectual content of this thesis is the product of my own work, except to the extent that assistance from others in the project's design and conception or in style, presentation and linguistic expression is acknowledged.'

19 December 2018

UNSW SYDNEY | Australia's Global University

Thesis/Dissertation Sheet

Surname/Family Name	:	**Noghrehchi**
Given Name/s	:	**Firouzeh**
Abbreviation for degree as give in the University calendar	:	**PhD**
Faculty	:	**Faculty of Science**
School	:	**School of Mathematics and Statistics**
Thesis Title	:	**Multiple imputation and access to likelihood-based tools in missing data problems**

Abstract 350 words maximum: (PLEASE TYPE)

Multiple imputation and maximum likelihood estimation (via the expectation-maximization algorithm) are two well-known methods readily used for analyzing data with missing values. While these two methods are often considered as being distinct from one another, multiple imputation (when using improper imputation) is actually equivalent to a stochastic expectation-maximization approximation to the likelihood. In this thesis we show how these two methods are equivalent, and further, exploit this result to show that familiar likelihood-based approaches can be used to enhance multiple imputation's performance in: (1) model selection, where familiar Akaike's Information Criterion (AIC) and the Bayesian Information Criterion (BIC) can be used to choose the imputation model that best fits the observed data; (2) hypothesis testing, where the familiar likelihood-ratio statistic can be used to perform composite hypothesis testing with multiple imputed data; (3) measurement error modelling, where familiar functional methods, such as Simulation-extrapolation and Corrected score, can be used to account for measurement error with multiple imputed data. We verify these results empirically and demonstrate the use of the methods on several classical missing data examples.

FOR OFFICE USE ONLY Date of completion of requirements for Award:

Abstract

Multiple imputation and maximum likelihood estimation (via the expectation-maximization algorithm) are two well-known methods readily used for analyzing data with missing values. While these two methods are often considered as being distinct from one another, multiple imputation (when using improper imputation) is actually equivalent to a stochastic expectation-maximization approximation to the likelihood. In this thesis we show how these two methods are equivalent, and further, exploit this result to show that familiar likelihood-based approaches can be used to enhance multiple imputation's performance in: (1) model selection, where familiar Akaike's Information Criterion (AIC) and the Bayesian Information Criterion (BIC) can be used to choose the imputation model that best fits the observed data; (2) hypothesis testing, where the familiar likelihood-ratio statistic can be used to perform composite hypothesis testing with multiple imputed data; (3) measurement error modelling, where familiar functional methods, such as Simulation-extrapolation and Corrected score, can be used to account for measurement error with multiple imputed data. We verify these results empirically and demonstrate the use of the methods on several classical missing data examples.

To Amir

Preface

This thesis is partially based on the following publications submitted to two peer-reviewed journals:

- Chapters 2 & 3 together are a modified version of Noghrehchi F., Stoklosa J., Penev S. and Warton D. I. (2018), "Imputation model selection for missing data: just use BIC!," *Statistics in Medicine*, in review.

- Chapter 5 is a modified version of Noghrehchi F., Stoklosa J. and Penev S. (2018), "Multiple imputation and functional methods in the presence of measurement error and missingness in explanatory variables," *Computational Statistics*, in review.

Acknowledgements

Thanks to my best friend and my partner, Amir, for his invaluable presence in my life and his never ending support. He watched me grow through the highs and the lows of these years, sometimes with joy, sometimes with pain, and offered me his hand or his shoulder or his ear or his sharp observation whenever I needed them. Without him, this thesis would not be finished.

To my co-supervisor, Dr Jakub Stoklosa, a great soul and a great supervisor, for his thorough support, extensive amount of time, precious comments, and his understanding.

To my supervisor, A/Prof Spiridon Penev, a great runner and a great teacher, for his genuine care, attention to the details, valuable advice, and his patience.

To my friend, Gordana, for her friendship, and all the fun.

To the staff of the School of Mathematics and Statistics, for making the school a better place to work.

To the blue groper and all the beautiful marine life in Gordons Bay, for making me forget my sadness.

Finally, thanks to Dr Andrew Weeks (The University of Melbourne), Richard Hill (Department of Environment, Land, Water and Planning Victoria) and the Department of Environment, Land, Water and Planning Victoria for collecting and providing the Eastern barred bandicoot data.

Contents

List of Figures

List of Tables

Abbreviation

The following table is a list of the most persistent abbreviation and the reader may find it useful for reference. However, abbreviation will be introduced in the text as needed.

AIC	Akaike Information Criterion
BIC	Bayesian Information Criterion
CS	Corrected score
EM	Expectation-maximization algorithm
MAR	Missing at random
MCAR	Missing completely at random
MCEM	Monte Carlo Expectation-maximization algorithm
MI	Multiple imputation
MLE	Maximum likelihood estimation
MNAR	Missing not at random
MSE	Mean squared error
LRT	Likelihood ratio test statistic
SIMEX	Simulation extrapolation
StEM	Stochastic Expectation-maximization algorithm

Notation

The following table is a list of the most persistent notation and the reader may find it useful for reference. However, notation will be introduced in the text as needed.

I	Identity matrix
$I(\theta)$	Fisher information matrix
$J(\theta)$	Observed information matrix
$KL(.\|\|.)$	Kullback-Leibler divergence
$l(\theta \mid y, r)$	Observed log-likelihood, also denoted by $l(\theta)$ for short when (y, r) is obvious
m	Number of missing observations
M	Number of multiple imputations
n	Sample size
$o(a_n)$	$b_n = o(a_n)$, for sequences a_n and b_n indexed by n, means $\forall \epsilon > 0$, $\exists n_0$ such that for $n > n_0$, $b_n/a_n < \epsilon$;
$O(a_n)$	$b_n = O(a_n)$ means $\exists C > 0$, C constant, and $\exists n_0$ such that for $n > n_0$, $b_n < Ca_n$;
$o_p(a_n)$	$X_n = o_p(a_n)$, X_n random variables indexed by n, means $X_n/a_n \to 0$ in probability;
$O_p(a_n)$	$X_n = O_p(a_n)$ means $\forall \epsilon > 0$ $\exists C > 0$ such that $P(\|X_n\| \leq Ca_n) \geq 1 - \epsilon$.
$p(. \mid .)$	Conditional distribution
r	Missingness indicator vector
$S(\theta)$	Score function
σ_U^2	Measurement error variance

T	Number of iterations of an iterative algorithm, at which the algorithm converges to its stationary distribution
θ	Model parameters
$\|\theta\|$	Number of model parameters
U	Measurement error
W	Error-contaminated explanatory variable
X	Explanatory variable
X_{mis}	Missing component of explanatory variable
X_{obs}	Observed component of explanatory variable
y	Observed data vector
Y	Response variable
z	Missing data vector

Chapter 1

Introduction

1.1 What are missing data?

"Can nothing be something?" When the Indians in the fifth century A.D. reached the idea of quantifying the absence of all quantity by zero that played a significant role in the history of mathematics. To brilliantly imagine a number of its own to denote absence revolutionalized the number system and gave rise to the fundamentals of mathematics we use today such as algebra and calculus (Kaplan, 2000).

In a similar spirit, Rubin (1976) proposed to recognize values that are not observed as "*missing*" values and introduced a random variable of its own to denote missingness. That is, we assume that if we had better techniques for data collection we would have observed their actual underlying values (Little and Rubin, 2002). Missingness is denoted by an indicator variable for whether or not a value is observed and, conditional on other variables in the dataset, comes from an underlying distribution *i.e.,* missingness mechanism, that captures the probabilistic reason for missing a value.

Missing data are planned observations not available for use in the analysis of a study. These values are missing from the intended sample of the population study after data collection and their presence in a data frame changes its shape away from a rectangular matrix, resulting in an "incomplete" data. Missing data are

surprisingly very common in quantitative research studies due to the limitations of the available techniques for data collection. Below, we look at some of the situations where missing data can appear, and further, discuss why the presence of missing data makes the statistical analysis special.

1.2 Where does missingness appear?

Missing data commonly arise in almost any quantitative research area. Examples include epidemiological, biological, agricultural, experimental and environmental studies as well as in social sciences, economics, psychology and criminology. Data can be missing on human subjects as well as on non-human ones, across time in longitudinal studies as well as cross-sectionally, confined to a single variable as well as to multiple variables, on response variables as well as on explanatory or auxiliary variables, for an entire unit of analyses as well as for single items measured on particular variables, from an interval on the unit of measurement as well as below or above a specified threshold, at individual levels as well as at group levels or at community levels. For example, suppose that we are interested in following up the mental health status of Australian adolescents in the state of New South Wales from secondary school year to their mid 20s in order to understand how mental disorders, such as depression and anxiety, persist into adulthoods. Furthermore, suppose that data were collected on a group that frequently moved their residential place and this posed difficulty in tracking some of the participants for at least once in adolescence as well as once in adulthood. For these participants who missed a whole wave (time point), data were missing for an entire unit of analysis (or unit nonresponse). Or, suppose that among the remaining participants in a particular wave, some of the depression scores (response) were missing because some of the more depressed participants were not motivated enough to answer the questionnaire. For others, some of the household incomes (explanatory) were missing due to sensitivity of the question. In these cases, data were missing for single items measured on particular variables (or item nonresponse).

Furthermore, missingness can occur when collecting data directly from the study subjects as well as when collecting data from existing records; at different stages

of a study *i.e.,* at the recruitment stage prior to the implementation of study, at the implementation stage, or at the follow-up stage in randomized clinical trials as well as in longitudinal studies (McKnight *et al.*, 2007, p. 5); can be related to the study subjects, the study design, and the interaction between the two (McKnight *et al.*, 2007, p. 5) as well as to administrative errors and measurement instruments' failures (Molenberghs and Kenward, 2007, p. 6). For example, suppose that we are interested in developing a housing price prediction model based on the property records of the past year in Darwin, Northern Territory of Australia. Furthermore, suppose that data were collected from the existing records of the individual real estate agencies. Since the property data are entered manually by the real estate agents, in a few records the listing or closing price was recorded as less than $1000, or either one or both prices were missing. In others, the total square meters of the house is recorded as smaller than the lot size, or the closing date is earlier than the listing date. These values are examples of missing data from existing records as well as missing data due to administrative errors.

Below, we discuss several motivating real-data examples in detail to elaborate further on where the missing data can appear. We will use the following examples as a basis for our numerical works in later chapters.

1.2.1 Motivating examples

We are motivated by three real-data examples in health research and by another two real-data examples in ecological research studies which consist of missing data. These datasets are quite different from each other varying in samples size, in the number of covariates and in response type. Furthermore, we are inspired by a food-safety research study where missingness occurs below a certain threshold and construct a simulation study based on these types of missingness later in Section 3.3.1. Below, we discuss these examples in details.

Pima Indian Women data. In our first example we analyze data collected on $n = 768$ Pima Indian women of at least 21 years of age living near Phoenix, Arizona who were tested for diabetes according to the World Health Organization

criteria. These data were collected by the US National Institute of Diabetes and Digestive and Kidney Diseases, and are available in the `mlbench` R-package. These data were analyzed in Miller *et al.* (1965) and further in Smith *et al.* (1988) and in Ripley (1996). The dichotomous response variable indicates whether or not diabetes was diagnosed within five years of the examination (positive/negative). The explanatory variables are

1. number of pregnancies,
2. plasma glucose concentration at 2 hours in an oral glucose tolerance test,
3. diastolic blood pressure (mm Hg),
4. triceps skin fold thickness (mm),
5. 2-hour serum insulin (μU/ml),
6. body mass index (kg/m^2),
7. diabetes pedigree function,
8. age (years).

Figure 1.1 shows the proportion of missingness for each pattern of multivariate missing data in the explanatory variables. There are 5 values missing (0.6%) for plasma glucose concentration, 11 values (1.4%) for body mass index, 35 values (4.6%) for diastolic blood pressure, 227 values (30%) for triceps and 374 values (49%) for serum insulin. Also, there are 192 values (25%) jointly missing for triceps and serum insulin. The relatively high proportion of joint missingness in triceps and serum insulin suggests that there might be a potential association between the variables that cause missingness in these two. Here, our interest is in predicting diabetes given all the explanatory variables whilst accounting for multivariate missing data in the predictors. In particular, we are not certain whether we should account for the missingness mechanism in the analysis or deem it to be ignorable. We address this problem in Section 3.6.

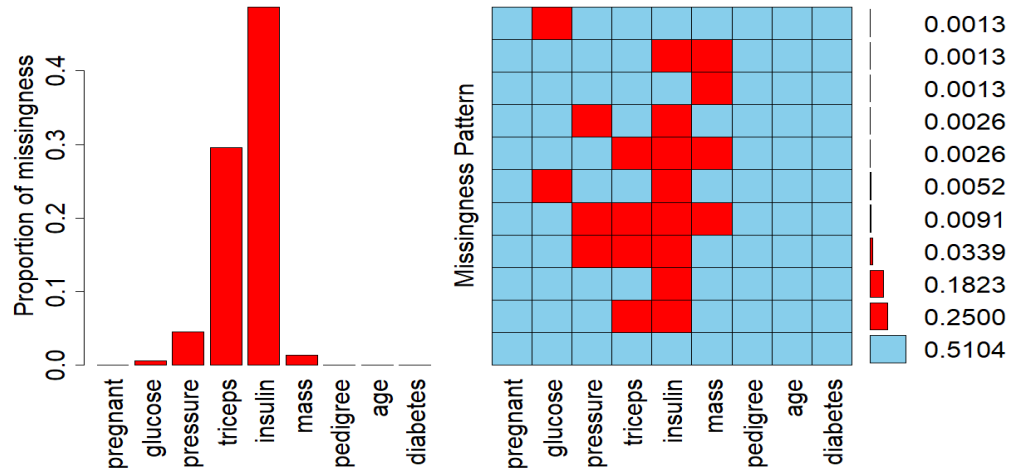


Figure 1.1. Missingness proportion (left) and missingness pattern (right) for the multivariate missing data in Pima Indian data. Most of the data is missing (in red) in serum insulin followed by triceps, where 49% of the data is missing in serum insulin and 30% in triceps, and with 25% of the data jointly missing in these two variables. About 51% of the data is fully observed (in blue).

Ozone data. In this example we analyze Los Angeles ozone pollution data collected on 366 observations in 1976 available in the `mlbench` R-package. This dataset was first analyzed in Breiman and Friedman (1985) and later in many studies such as in Buja *et al.* (1989) and in Friedman and Silverman (1989), and more recently, in Eugster and Leisch (2011) and in Wang *et al.* (2015). It consists of a response variable – the daily maximum one-hour average *ozone reading*, and of 9 meteorological explanatory variables. The meteorological variables are

1. temperature (degrees F) measured at
   - El Monte, CA
   - Sandburg, CA
2. inversion base height (feet) measured at Los Angeles International Airport (LAX)
3. humidity (%) measured at LAX
4. visibility (miles) measured at LAX
5. wind speed (mph) measured at LAX

6. pressure gradient (mm Hg) from LAX to Daggett, CA

7. inversion base temperature (degrees F) at LAX

8. 500 millibar pressure height (m) measured at Vandenberg AFB

There are 139 missing observations for temperature recordings at El Monte, indicating about 38% missing proportion in this variable. Measurements of climate variables (such as, temperature, humidity, rainfall, etc.) are usually susceptible to some uncertainty, see Stoklosa *et al.* (2015) for further details and references. We therefore assumed that temperature is subject to classical measurement error (see Section 1.5.3). Different studies have shown that temperature is the most influential predictor of ozone reading (Breiman and Friedman, 1985; Efron and Tibshirani, 1986; Casella and Moreno, 2006; Miller, 2002). Here, for our purposes, we are interested in predicting ozone reading given temperature at El Monte, inversion base height and humidity whilst accounting for both missingness and measurement error in the temperature variable. We address this problem in Section 5.4.

Survival of infants data. The following example is taken from Example 9.8 of Little and Rubin (2002) where a 2^3 contingency table on the survival of infants is analyzed in the presence of missing data. Suppose we have three dichotomous variables of Prenatal care ($i = \{Less, More\}$), Survival ($j = \{Died, Survived\}$), and Clinic ($k = \{A, B\}$).

Let n_{ijk} denote the total count of the ijk^{th} cell, and $n = \sum_i \sum_j \sum_k n_{ijk}$ be the total sample size. Table 1.1 is (artificially) partially classified for about 26% of the data (255 cases out of n =970) where we observed only the sums $n_{ij.}$ instead of all n_{ijk} values. The remainder of the data (715 cases out of 970) are completely observed. Also, let π_{ijk} denote classification probability for the ijk^{th} cell. The response variable is the cell counts, n_{ijk}. Subject to missingness, these cell counts can be modelled by assuming different associations between the three predictors: Survival (S), Prenatal care (P), and Clinic (C). Here, our objectives are (1) to select the best association between predictors that explains the missingness in the

cell counts and (2) to select the best association between predictors that explains the data whilst accounting for the missing cell counts. We address both of these problems in Sections 3.4 and 4.4, respectively.

Table 1.1. Survival of infants taken directly from Example 9.8 of Little and Rubin (2002).

Clinic (C)	Prenatal care (P)	Survival (S)		
		Died	*Survived*	
A	*Less*	3	176	
	More	4	293	
B	*Less*	17	197	
	More	2	23	complete = 715
?	*Less*	10	150	
	More	5	90	partial = 255

Left-censored data. Nitrate is a chemical element that is part of the nitrogen cycle in nature and can be found in soil, water, and biomass. In addition, high levels of nitrate can also be found in plants and vegetables due to its significant role as a fertilizer in agriculture. However, consumption of high amounts of nitrates from vegetables as well as inappropriate storage of cooked vegetables can potentially lead to adverse health effects such as methaemoglobinemia and carcinogenesis (EFSA, 2008, 2010). Quijano *et al.* (2017) studied a total of 533 samples of seven vegetable species in order to investigate the toxicological risk associated with the intake of vegetables exposed to nitrate. The vegetable species included carrot, lettuce, ice-berg lettuce, artichoke, chard, spinach and potato. Data on nitrate levels were collected from the Valencia Region, Spain from 2009 to 2013. However, between $0\% - 62.5\%$ of the measured nitrate concentration levels were missing below a limit of detection $(80\ \mathrm{mg/kg})$ among the seven vegetable species. Observations that are missing below the limit of detection of the instrument or technique used to measure observations are referred to as *left-censored* in the literature. The terminology of left censoring emerges from the fact that the left side of the distribution of sample values is truncated at the limit of detection. Left-censoring commonly occurs in agricultural, environmental, epidemiological, biological and occupational studies. We did not have access to this dataset, but we used it as an inspiration for a simulation study with left-censored values in Section 3.3.1.

Eastern barred bandicoot data. In this example we analyze data on $n = 77$ Eastern barred bandicoots *Perameles gunnii* in Hamilton, South-eastern Victoria, Australia collected in November, 2012. We considered two covariates that were collected during trapping: *gender*, which was correctly identified each time an individual was seen; and *body weight* (*weight)*, which was missing on some occasions upon capture. There were 50 unique females and 27 males captured in this study period. There were 14 individuals without the record of body weight, hence the proportion of missing data was 18.2%. The same dataset was used in Stoklosa *et al.* (2019) where the interest was in estimating the population size of bandicoots via capture-recapture methods. Here, our interest is in predicting body weight given the bandicoot gender whilst accounting for missing data in the body weights. However, we are not certain whether we should account for the missingness mechanism in the analysis or deem it to be ignorable. We address this problem in Section 3.5.

1.3 Why does missigness make statistical analysis special?

Missing data hinders the ability to apply standard statistical analyses designed for complete datasets (Schafer and Graham, 2002). The presence of missing data in a dataset makes the data matrix to lose its shape as a rectangular matrix. Since most of the standard statistical analyses were designed for handling data in rectangular matrix shapes, *e.g.,* regression analysis, missing data makes the data unsuitable for standard statistical analyses and complicates the analysis.

Moreover, missing data can restrict the ability to draw final conclusions from a study and could lead to incorrect inferences (Graham, 2009). Missing data, in particular in medium to high proportions, can affect the conclusions from a study and will compromise their validity unless missingness is completely due to chance, a strong assumption that is not always met. This is because the study aimed to make inferences about some aspects of an intended sample which would be fully observed if we had better techniques for collecting data. More specifically, missing

data can lead to incorrect inferences by (1) producing bias in parameter estimation (Rubin, 1987; Schafer, 1997; Little and Rubin, 2002, p. 19) and (2) resulting in efficiency loss (Little and Rhemtulla, 2013; Little and Rubin, 2002, p. 19). Missing data can lead to biased estimates of parameters if they systematically differ from observed data in terms of one or more key variables (Raghunathan, 2004), that is, if they reduce representiveness of the samples for analysis. Bias hinders the ability to generalize the results of a study, since the result would be different if we had observed the missing values. Furthermore, missing data can cause efficiency loss and increased standard errors if they carry some information about the parameters that is not captured in the observed values. The magnitude of efficiency loss depends on the proportion of missing information due to missingness as well as on the objective of the analysis (Carpenter and Kenward, 2012, p. 9).

Finally, it should also be pointed out that missing data can decrease statistical power of a study (Peng *et al.*, 2006). Statistical power is the probability of rejecting the null hypothesis when is it false. Unless the number of observed data is still substantially large, the lost data will result in fewer observations available for analysis and this reduction in sample size will reduce statistical power to detect a significant effect depending on the proximity of the effect size to the null value.

1.4 Structure of the remaining chapters

In this thesis, we establish the close link between two different missing data analysis approaches commonly used in the literature to handle missingness, namely multiple imputation and maximum likelihood estimation. This connection is further exploited by showing that standard likelihood-based tools available in the maximum likelihood estimation literature are applicable to problems in the multiple imputation literature. Specific contributions contained in this thesis focus on addressing the question of: (1) selecting an imputation model to use to impute missing values with multiple imputation, (2) hypothesis testing and testing of goodness-of-fit with multiple imputation, and (3) measurement error modelling with multiple imputed data.

The following section 1.5 in this chapter introduces some topics relevant to the remaining chapters and is included for general reference. The rest of the thesis is organized as follows. Chapter 2 looks at two well-known missing data analysis methods in the literature, namely multiple imputation and stochastic expectation-maximization algorithm as stochastic approximation to maximum likelihood estimation, and shows how these two methods are equivalent. Chapters 3–5 explore some of the contexts in which this equivalence allows access to likelihood-based tools for enhancing multiple imputation's performance. Chapter 3 investigates access to likelihood-based tools such as information criteria for imputation model selection. Chapter 4 investigates access to likelihood-based tools such as likelihood ratio statistic for hypothesis testing and testing of model goodness-of-fit with multiple imputed data. Chapter 5 numerically investigates access to likelihood-based tools such as Simulation extrapolation and Corrected score to be combined with multiple imputation in order to account for the combined effect of missingness and measurement error in explanatory variables. Finally, Chapter 6 provides a summary discussion and outlines avenues for future research.

1.5 Toolbox for the remaining chapters

In this section we provide a brief outline of some well-known methods which will be used in various chapters throughout the thesis. The methods presented in this toolbox are not new and are simply intended to give the reader a clearer exposition and quick referral of the methods used. We will not review standard statistical methodology like Akaike's Information Criterion and likelihood ratio tests but will refer the reader to various citations if they appear throughout this thesis. All coding in this thesis was done in R (R Development Core Team, 2019).

1.5.1 Missing data analysis

Suppose y is a vector of observed data, z is a vector of the missing data, and r is a vector of missingness indicators where each component of r is either 0 if the data point is missing or 1 if the data point is observed. Also, θ denotes a vector of unknown model parameters. For now we assume that these data are missing

not at random (MNAR) but note that the methods presented below are also applicable for missing at random (MAR) data, since this is a special case of MNAR where missingness is ignorable, see Rubin (1976) and Little and Rubin (2002, pp. 11–12) for a detailed discussion on the types of missingness mechanisms. Data are MAR if missingness depends only on the observed variable whereas data are MNAR if missingness depends on the observed variable as well as the missing variable itself. Missingnes is ignorable if the observed variable is considered to be sufficient to offset the effects of missingness *i.e.,* data are MAR. Under the ignorability assumption, the conditional distribution of the missing variable given the observed variable and the missingness indicator, $p(z \mid y, r, \theta)$, will not depend on the missingness model anymore and can be simplified as $p(z \mid y, \theta)$.

Let $p(y, r \mid \theta)$ be the observed data likelihood and $p(y, z, r \mid \theta)$ be the complete data likelihood. To estimate θ, we maximize the observed data likelihood, $p(y, r \mid \theta)$, which, in the presence of missing data, is obtained by integrating out the missing data from the complete-data likelihood, $p(y, z, r \mid \theta)$:

$$L(\theta; y, r) = p(y, r \mid \theta) = \int p(y, z, r \mid \theta) \mathrm{d}z. \tag{1.1}$$

However, the likelihood (1.1) may not be available in a closed form because of analytically intractable integration or can be difficult to solve for many situations in practice, such as for complex models or when the data are of high dimension. We point out that in (1.1) and further in the thesis, we will avoid specifying the region of integration when there is no confusion arising by doing that.

A common approach to missing data problems is to impute (fill-in) missing data, z, with some plausible values that are a good summary of z, to which we obtain a (pseudo-) complete dataset. The rationale underlying imputation of the missing values is to edit an incomplete dataset into a complete dataset (rectangular matrix) in order to be able to apply standard statistical analysis methods. The goal of imputation is to combine available information from the observed data with statistical assumptions about the missingness mechanism in order to obtain valid inferences about a population (Dong and Peng, 2013). Note that, the goal of imputation is to obtain valid inferences from the data rather than estimation of the

missing values. In Chapter 2, we discuss two well-known missing data analysis methods that use (multiple) imputation to account for missingness in the analysis.

1.5.2 Variational approximation

Variational approximation (Jordan *et al.*, 1999) refers to a class of deterministic approximation techniques that are based on variational methods. The root of variational methods lies in the "calculus of variations", hence the terminology. Calculus of variations is a field of mathematical analysis that seeks to optimize a functional over a class of functions on which that functional depends, using variations (small changes) in the functions and the functional itself. Although there are no approximations in the variational theory, variational methods can be used to find approximate solutions in statistical inference and estimation when some restriction is imposed on the class of functions, usually in a way to enhance tractability (Ormerod and Wand, 2010). Variational methods as approximation techniques are readily applied in a wide range of settings, including regression models with missing data (Faes *et al.*, 2011), hidden Markov models (Foti *et al.*, 2014), time series models (Archambeau *et al,*, 2007), hierarchical models (Woolrich *et al.*, 2004), and Gaussian process models (Hensman *et al.*, 2013).

The key idea underlying variational approximation is to allow for an optimization problem to be relaxed by approximating the function to be optimized. Variational approximation can be defined as a (trade-off) method for optimizing a likelihood function while enhancing its tractability, hence, making approximate inference for model parameters. Variational approximation approximates a likelihood function by other likelihood functions for which inference is more tractable while the approximations are guided by a discrepancy measure. Ormerod and Wand (2010) and Blei *et al.* (2017) gave detailed explanations of variational approximation using familiar examples for statisticians and pointed out to some relevant literature on variational approximation.

We make use of variational approximation techniques in the missing data context. We briefly discuss variational approximation here and see its use in imputation

model selection in Chapter 3.

One particular situation where variational approximation is useful is when the likelihood specification involves integrating out a latent variable z, as given in equation (1.1). Suppose that our interest is in obtaining $\hat{\theta} = \arg\max_\theta \log p(y, r \mid \theta)$ where

$$\log p(y, r \mid \theta) = \log \int p(y, z, r \mid \theta) \mathrm{d}z,$$

however, evaluating $\log p(y, r \mid \theta)$ by a tractable marginalizing over the latent variable z is difficult. Our motivation for using variational approximation in the missing data context stems from this.

Denote $p(z \mid y, r, \theta)$ as the exact conditional distribution of the latent variable z given the observed variables (y, r) and let $q(z)$ be a variational distribution over the latent variable z. Define an approximate function Q, for which local optimization is computationally easier, as

$$Q = -KL(q||p) + \log p(y, r \mid \theta),$$

where $KL(.||.)$ denotes the Kullback–Leibler divergence (Kullback and Leibler, 1951) and $KL(q||p)$ measures the discrepancy of the arbitrary approximation $q(z)$ of the likelihood from the exact likelihood $p(z \mid y, r, \theta)$. Kullback–Leibler divergence is given by

$$KL(q||p) = \int q(z) \log \frac{q(z)}{p(z \mid y, r, \theta)} \mathrm{d}z.$$

The approximate function Q has the following characteristics:

1. $\forall q(z)$, we have $Q \leq \log p(y, r \mid \theta)$ *i.e.,* Q is the *lower bound* for $\log p(y, r \mid \theta)$ over z

2. the slack in the bound is given by $KL(q||p)$, more specifically:

   - if $KL(q||p) = 0$ then $Q = \log p(y, r \mid \theta)$ *i.e.,* the exact observed log-likelihood function $\log p(y, r \mid \theta)$ is recovered when there is no divergence between the variational likelihood function $q(z)$ and the exact likelihood function $p(z \mid y, r, \theta)$, that is, when $q(z) = p(z \mid y, r, \theta)$

- for $KL(q||p) > 0$, the larger $KL(q||p)$ the more $q(z)$ diverges from the exact likelihood $p(z \mid y, r, \theta)$.

We approximate the log-likelihood function $\log p(y, r \mid \theta)$ by a $q(z)$ for which a lower bound Q is more tractable than $\log p(y, r \mid \theta)$ and obtain a variational approximation to $\arg\max_\theta \log p(y, r \mid \theta)$ by solving a new maximization problem over the lower bound $\arg\max_\theta Q$, where tractability is achieved by restricting $q(z)$ to a more manageable class of distributions. Clearly, maximization of the lower bound Q is equivalent to minimization of the Kullback–Leibler divergence of $q(z)$ from $p(z \mid y, r, \theta)$.

1.5.3 Measurement error modelling

Measurement error (or errors-in-variables) modelling is a well-known technique used to correct for measurement uncertainty in explanatory (or predictor) variables (Carroll *et al.*, 2006). As discussed in the Ozone data example in Section 1.2.1, temperature variables in the study were subject to uncertainty in the measurement and consisted of missing values. We give a brief outline and some generic notation used in measurement error modelling, we then present some new methods which incorporate missing data into this framework in Chapter 5.

Consider a random variable X which is usually the predictor variable for a regression model. We suppose that X is subject to zero-mean measurement error U when we are unable to observe it directly and instead we observe an error-contaminated measurement for X, which we denote as W. Let σ_U^2 be the measurement error variance associated with U. This quantity is usually assumed known but can also be estimated from repeated measurements of X or from validation data (if available). Measurement errors can have two underlying structures for relating W to X (Carroll *et al.*, 2006, p. 26–32):

- *classical measurement error* models, which model the conditional distribution of W given X. This model assumes that $W = X + U$, where U is the measurement error, a random variable with mean 0 and variance $\sigma_U^2 > 0$ and is independent from X. This model assumes that W is an unbiased measure-

ment of X, and that W has larger variability than X.

- *Berkson measurement error* models, which model the conditional distribution of X given W. This model assumes that $X = W + U$, where U is the measurement error with mean 0 and variance $\sigma_U^2 > 0$ and is independent from W. This model assumes that X has larger variability than W.

The classical measurement error is identified through the assumption that a measurement error is independent from the true unobserved variable. Otherwise, the measurement error is a Berkson error. Classical measurement errors are more commonly studied in the literature due to the wide range of examples associated with this type of error. The implication of a classical measurement error is perhaps easily demonstrated in a simple linear regression analysis with explanatory variable subject to measurement error. Suppose that $Y = (Y_1, \ldots, Y_n)^\intercal$ is the response variable and we are interested in the following linear model

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \quad \text{with } \mathrm{E}(\epsilon_i) = 0,\ \mathrm{Var}(\epsilon_i) = \sigma_\epsilon^2, \tag{1.2}$$

$i = 1, \ldots, n$, where X_i is observed with an additive, independent measurement error,

$$W_i = X_i + U_i, \quad \text{with } \mathrm{E}(U_i) = 0,\ \mathrm{Var}(U_i) = \sigma_U^2. \tag{1.3}$$

From (1.2) and (1.3), the regression of Y on X may be written as

$$Y_i = \beta_0 + \beta_1 W_i + \eta_i, \quad \eta_i = \epsilon_i - \beta_1 U_i. \tag{1.4}$$

This model resembles an usual simple linear regression model except that the error term is clearly dependent on β_1. Let $\hat{\beta}_{Y|X}$ and $\hat{\beta}_{Y|W}$ be the least square estimators of the slope coefficient β_1 from regression of Y on X in (1.3) and from regression of Y on W in (1.4), respectively. Since W_i and U_i are correlated with $\mathrm{cov}(W_i, U_i) = -\beta_1 \sigma_U^2$, it can be shown that

$$\hat{\beta}_{Y|W} = \hat{\lambda}\ \hat{\beta}_{Y|X} + o_p(1), \quad \hat{\lambda} = \frac{s_X^2}{s_X^2 + s_W^2},$$

where s_X^2 and s_W^2 denote the sample variances of X and W, respectively (Stefanski,

2000). Hence, ignoring the measurement error will result in bias attenuation (bias towards 0) in the slope estimator due to the fact that $0 < \hat{\lambda} < 1$. This result can easily be extended to multiple linear regression models. For more technical details in linear models see Fuller (1987) and Cheng and Van Ness (1999). For technical details in nonlinear models see Carroll *et al.* (2006), and in nonparametric models see Hall *et al.* (2018).

Besides error structure, the properties of the unknown values of X is a defining characteristic in the measurement error analysis (Carroll *et al.*, 2006, p. 25). Data structure of X can be modelled by

- *functional modelling* methods, where X can be either fixed or random with no distributional assumptions, or by
- *structural modelling* methods, where X is random with distributional assumptions (usually Gaussian).

In this thesis, we strictly focus on classical measurement error structures in explanatory variables. Various methods have been developed to handle measurement error in explanatory variables, see Carroll *et al.* (2006). In Chapter 5 we review two of these methods but focus on functional likelihood-based approaches known as *Simulation extrapolation* (Cook and Stefanski, 1994) and *Corrected score* (Nakamura, 1990) which we use within a missing data framework.

1.5.4 Metropolis-within-Gibbs sampler

Metropolis-within-Gibbs sampler is a hybrid of two well-known Markov chain Monte Carlo (MCMC) methods, namely the Metropolis algorithm (Metropolis and Ulam, 1949; Metropolis *et al.*, 1953) and Gibbs sampling (Geman and Geman, 1984; Gelfand and Smith, 1984). This is a powerful method used in various applications which allows for fitting sophisticated models. More specifically, MCMC methods are useful in situations where it is difficult to sample directly from a target distribution, such as in missing data problems, as in equation (1.1).

A Markov Chain is a sequence of steps, in which the next step of the chain depends

only on the current step so that we may forget the past. Under certain regularity conditions, the chain, with each step, gets closer to stabilising at a unique probability distribution, called the stationary or equilibrium distribution (Gelman *et al.*, 2013b, p. 275). Monte Carlo is a general term for a simulation technique based on Monte Carlo integration. The Monte Carlo component is responsible for sampling from approximate distributions at each step and ensures that the random draws of missing data are independent (Schafer, 1999). A thorough discussion on MCMC methods is given in Tanner and Wong (1987).

We will apply the Metropolis-within-Gibbs sampler in our numerical works in Sections 3.2.4 and 3.3.2 to simulate values for missing data in multivariate missing data problems. We therefore provide the reader with some brief details here.

Metropolis algorithm. Suppose our goal is to generate samples from some distribution $q(z)$ where $q(z) = f(z)/c$, and the positive normalizing constant c is not known or is very difficult to compute. The Metropolis algorithm generates a sequence of samples from this distribution as follows. The algorithm starts from some arbitrary initial value z^0 that satisfies $f(z^0) > 0$. Given current value of z',

(a) sample a value z'' from some proposal distribution $g(z'' \mid z')$, which is symmetric and denotes the probability of returning a value of z'' given a current value of z',

(b) accept z'' with probability $\min\left\{\dfrac{f(z'')}{f(z')}, 1\right\}$, otherwise retain z'.

The sequence $(z^0, z', z'', \ldots)$ generates a Markov chain and when the steps (a)–(b) are repeated a sufficiently large number of times, the chain eventually provides samples from $q(z)$, which do not depend on the starting values. The choice of a symmetric proposal distribution can, for example, be based on a random walk chain with $z' = z'' + \epsilon$, where $\epsilon \sim N(0, \sigma_\epsilon)$. In this context, σ_ϵ is referred to as the tunning parameter and can be adjusted so that the chains are mixed well *i.e.,* the space is explored.

Gibbs sampling. Suppose our goal is to generate samples from some joint distribution $p(z_1, \ldots, z_k)$, where it is far easier to sample from a sequence of univariate conditional distributions, rather than to obtain the marginals by integration of the joint density. The Gibbs sampler draws from univariate conditional distributions in an iterative manner to eventually sample from the distribution of a multivariate variable and proceeds as follows. Starting from some arbitrary initial values, at iteration t, generate a multivariate sample $(z_1^{(t)}, \ldots, z_k^{(t)})$ by randomly drawing the random variable z_j, $j = 1, \ldots, k$, in a successive manner, from its univariate conditional distribution given the current values of all the other variables:

$$
\begin{aligned}
z_1^{(t)} &\sim p(z_1 \mid z_2^{(t-1)}, \ldots, z_k^{(t-1)}) \\
&\vdots \\
z_j^{(t)} &\sim p(z_j \mid z_1^{(t)}, \ldots, z_{j-1}^{(t)}, z_{j+1}^{(t-1)}, \ldots, z_k^{(t-1)}) \\
&\vdots \\
z_k^{(t)} &\sim p(z_k \mid z_1^{(t)}, \ldots, z_{k-1}^{(t)}).
\end{aligned}
$$

This process is iterated for a sufficiently large number of times until the Gibbs sampler eventually provides a sample from $p(z_1, \ldots, z_k)$, which would not depend on the starting values.

Metropolis-within-Gibbs sampler. The Metropolis-within-Gibbs sampler aims to generate samples from a multivariate distribution by successively sampling from associated univariate conditional distributions (the Gibbs part) and by using one Metropolis step instead of a direct sampling from each conditional distribution (the Metropolis part) (Gamerman and Lopes, 2006, p. 211-214). This algorithm performs Metropolis within each iteration of Gibbs sampling as follows. Suppose our goal is to generate samples from $p(z_1, \ldots, z_k \mid y, r, \theta)$, where y and r are vectors of size $(k \times 1)$. Furthermore, suppose that a current approximation to the univariate conditional distribution of z_j, $j = 1, \ldots, k$, up to a normalizing constant, is known:

$$
p(z_j \mid z_1, \ldots, z_{j-1}, z_{j+1}, \ldots, z_k, y_j, r_j, \theta) \propto p(z_1, \ldots, z_k, y_j, r_j \mid \theta).
$$

At iteration t of the Metropolis-within-Gibbs sampler, for each j^{th} component of z, in a successive manner we

(a) candidate a value z'_j by randomly drawing from a proposal distribution $g(z'_j \mid z_j^{(t-1)})$ that is symmetric, such as a normal distribution,

$$z'_j \sim g(z'_j \mid z_j^{(t-1)})$$

(b) accept $z_j^{(t)} = z'_j$ with probability

$$\min\left\{\frac{p(z_1^{(t)},\ldots,z_{j-1}^{(t)},z'_j,z_{j+1}^{(t-1)},\ldots,z_k^{(t-1)},y_j,r_j \mid \theta^{(t-1)})}{p(z_1^{(t)},\ldots,z_{j-1}^{(t)},z_j^{(t-1)},z_{j+1}^{(t-1)},\ldots,z_k^{(t-1)},y_j,r_j \mid \theta^{(t-1)})},1\right\},$$

otherwise keep $z_j^{(t)} = z_j^{(t-1)}$.

In the same way as with the original Gibbs sampler, this process is iterated for a sufficiently large number of times until the algorithm eventually provides a sample from $p(z_1,\ldots,z_k)$, which does not depend on the starting values. For more technical details see Tierney (1994), Gilks *et al.* (1995) and Gamerman and Lopes (2006).

Chapter 2

Multiple imputation and access to likelihood-based tools

2.1 Background

Multiple imputation (MI, Rubin, 1987) and maximum likelihood estimation (MLE) via the expectation-maximization algorithm (EM, Dempster *et al.*, 1977) are perhaps the most well-known methods used to account for missing values in partially observed datasets. MI is a Monte Carlo approach developed for handling missing data in sample surveys where missing values are imputed to create a complete dataset, whereas the EM algorithm is an iterative method developed for finding the MLE in the presence of missing data. Stochastic versions of EM, in particular Monte Carlo EM (MCEM, Wei and Tanner, 1990) and stochastic EM (StEM, Broniatowski *et al.*, 1983; Celeux and Diebolt, 1985, 1987), are numerical approaches of EM to compute a Monte Carlo approximation to MLE. Due to the differences in the computation mechanics and several theoretical properties, MI and MLE are often recognized as two distinct approaches in the missing data literature.

Surprisingly, the connection between MI and MLE has been seldom addressed in the literature. There exists a type of MI, *improper MI*, which does not result in statistically valid inference when based on Rubin (1987)'s combination rules of MI. Wang and Robins (1998) identified a class of stochastic versions of EM as improper (or obviation from full Bayesian) MI, and compared their asymp-

totic properties with MI's. von Hippel (2012) used these results and developed a simpler variance estimator for non-iterative StEM based on within- and between-imputation variances defined in Rubin (1987)'s combination rules. Also, Biscarat *et al.* (1992) stated that MI can be understood as a Bayesian version of StEM, and StEM can be recovered from an MI algorithm with non-informative priors. Additionally, Wei and Tanner (1990) proposed MCEM and compared it to a type of MI (Data Augmentation), and pointed out their difference in the computation of posterior distribution. So far, the main emphases have been on studying the differences between MI and stochastic versions of EM, rather than on their connection.

In this chapter, we exploit the connection between MI and StEM and discuss how these two methods are equivalent. The importance of this connection is in the way it changes our point of view towards the relationship between these two methods. This new point of view, which is the novel contribution of this chapter, allows ideas to move between the two literatures in order to enhance and improve these methods' performances. Specifically, due to a range of available likelihood-based tools and their desirable statistical properties, MI's performance could be improved by borrowing ideas from the maximum likelihood literature and accessing likelihood-based tools.

The outline of this chapter is as follows. We begin by giving details of both methods, specifically focusing on some theoretical properties, obtaining standard errors via combinations rules and discussing how to choose the number of imputations. Although choosing the number of imputations is not the focus of this thesis, it is important to know how the asymptotic properties of MI and StEM estimators compare for finite number of imputations, and further, how the number of imputations is chosen in both the MI and the ML literature. These details are given as literature reviews on MI in Section 2.2 and on StEM in Section 2.3. Finally, we briefly discuss a few areas where we could gain from this connection. We will study these potential gains in more detail in later chapters. Also, we reserve applications to real-data until later chapters. The novel contributions of this chapter are drawing on the close connection between MI and StEM in Section 2.4 and the potential gains from their connection discussed in Section 2.5.

2.2 Multiple imputation

2.2.1 Definition

Multiple imputation is a Monte Carlo method originally proposed by Rubin (1976, 1987) where every missing observation in the dataset is imputed with a simulated value to create a complete dataset. The imputation step is repeated $M \geq 2$ times, such that M-completed datasets are generated. That is, we impute z with a set of plausible values, $z^* = (z^{(1)}, z^{(2)}, \ldots, z^{(M)})$, for some $M \geq 2$. Each (pseudo-) complete dataset is then separately analyzed by standard complete-data analysis methods. In order to achieve asymptotically valid statistical inference, the resulting estimates from each M-completed dataset need to be pooled according to Rubin's rule (Rubin, 1987) as discussed in this chapter (see Section 2.2.2).

Rubin (1976, 1987) proposed MI in response to the shortcomings of single imputation methods when dealing with nonresponse in surveys and provided a thorough justification of the method. The main difference between MI and single imputation lies in the way they treat imputed data. Inferences based on single imputation treat imputed data as if they were true observed data whereas in reality we do not have as much certainty in the imputed values as we would have were they observed. Inferences based on MI allows us to reflect the uncertainty in the missing data by combining the results of the multiple imputations. However, MI is not restricted to survey analysis and can also be used to impute missing data in any setting.

The foundation for creating MI arises from a Bayesian framework. The work needed to create multiple imputations can be divided into three tasks, which consist of the modelling, the estimation, and the imputation task (Rubin, 1987). The modelling task chooses a Bayesian model for the complete data and missingness mechanism. The estimation task formulates the posterior distribution of the parameters of the chosen model. The imputation task draws a value from the posterior parameter distribution, and given this value takes a random draw from the posterior predictive missing data distribution.

Estimation for θ is carried out as follows. Consider a complete dataset (y, z, r), the complete-data model $p(y, z, r \mid \theta)$, and the posterior predictive distribution of missing data given the observed data $p(z \mid y, r)$. Multiple imputations are repeated random draws from $p(z \mid y, r)$. The Monte Carlo average of the completed-data posterior distributions gives an approximation to the observed posterior,

$$p(\theta \mid y, r) = \int p(\theta \mid y, z, r) p(z \mid y, r) \mathrm{d}z \simeq M^{-1} \sum_{j=1}^{M} p(\theta \mid y, z^{(j)}, r). \tag{2.1}$$

In practice, we often do not know $p(z \mid y, r)$, which in turn, can be approximated based on its relationship to the observed posterior of θ,

$$p(z \mid y, r) = \int p(z \mid y, r, \theta) p(\theta \mid y, r) \mathrm{d}\theta$$

where $p(z \mid y, r, \theta)$ is the predictive distribution of missing data given the observed data and a current estimate of θ. We refer to $p(z \mid y, r, \theta)$ as the *imputation model* throughout this chapter. Therefore, MI is commonly performed in an iterative manner to approximate the observed posterior of θ as outlined below.

In order to simulate from the posterior distribution of θ, we first draw missing values from the current approximation to the imputation model, $p(z \mid y, r, \theta^{(0)})$, based on some initial guess for θ, and use the drawn values to complete the dataset. Then, we draw $\theta^{(t)}$, $t = 1, \ldots, T$, from its completed-data posterior distribution given the current imputed values $p(\theta \mid y, z^{(t-1)}, r)$. These two steps of imputation (I-step) and posterior estimation (P-step) are iterated a large number of times ($t = T$) to eventually produce a draw of (z, θ) from their joint observed posterior. Finally, the next M imputations are implemented as multiple imputations in Equation (2.1) to approximate the posterior of θ by averaging over repeated draws of θ. These steps can be summarized in the following algorithm:

MI algorithm

0. Fix $\theta^{(t)}$ in Θ, $t = 0$
1. Draw $z^{(t+1)}$ from $p(z \mid y, r, \theta^{(t)})$
2. Draw $\theta^{(t+1)}$ from $p(\theta \mid y, r, z^{(t+1)})$
3. Set $t = t + 1$ and repeat steps 1–2 until convergence at $t = T$

4. Repeat steps 1–2 for extra $M \geq 2$ iterations, $t = T+1, \ldots, T+M$, to create multiple imputations
5. Combine the results of the final M iterations by Rubin's rule

Iterative MI can be understood as a special case of Data Augmentation (Tanner and Wong, 1987). Data Augmentation is an iterative method developed for approximating the posterior distribution in the presence of missing data, where in the I-step, $M \geq 1$ imputations are created (*augmented*) and in the P-step the posterior of θ is updated as the mixture of the M-completed posteriors. The value for M can change at each iteration of Data Augmentation. Although this thesis focuses on the case where the missing data are imputed only once in the I-step of MI until the algorithm converges, if using Data Augmentations, it is possible to run MI by setting $M > 1$ in the I-step. It is worth noting that Data Augmentation was motivated by the EM algorithm offering a Bayesian alternative for situations when statistical inference based on MLE and the associated standard error cannot be reliable because the likelihood cannot be approximated accurately. Analogous to Data Augmentation, MI approximates posterior distributions.

Equation (2.1) substantially simplifies the process of obtaining the posterior distribution of θ by enabling us to draw from a combination of two simpler posteriors. It follows that by using simulated values of z, the posterior mean and variance of θ can similarly be approximated, for large M. This result is used in Rubin's combination rules to make inference about the model parameters. Note that MI's three tasks can create multiple imputed datasets in trivial situations by doing explicit computations provided, however, in nontrivial situations more sophisticated computational techniques such as MCMC are required.

2.2.2 Combination rules

MI's combination rules (or Rubin's rule) are derived based on the Bayesian framework where θ and its estimate $\hat{\theta}$ are treated as unobserved random variables where $\hat{\theta}$ denotes an estimate of θ in the absence of missing data. Further, with complete data, inferences about θ would be based on a normal approximation assumption $(\theta - \hat{\theta}) \sim N(0, \mathcal{W})$ where $\mathcal{W}$ is the associated variance of $(\theta - \hat{\theta})$. In the presence of

missing data, the mean and variance of the posterior distribution of θ are given by

$$\mathrm{E}(\theta \mid y, r) = \mathrm{E}\big[\mathrm{E}(\theta \mid y, r, z) \mid y, r\big]$$

and

$$\mathrm{Var}(\theta \mid y, r) = \mathrm{E}\big[\mathrm{Var}(\theta \mid y, r, z) \mid y, r\big] + \mathrm{Var}\big[\mathrm{E}(\theta \mid y, r, z) \mid y, r\big], \tag{2.2}$$

respectively. Based on this result, in the presence of missing data, the posterior mean and variance of θ can be approximated as described below.

Suppose that under a particular Bayesian model, $\hat{\theta}_1, \ldots, \hat{\theta}_j$ and $\mathcal{W}_1, \ldots, \mathcal{W}_j$ are the obtained values of $\hat{\theta}$ and $\mathcal{W}$ for each of $j = 1, \ldots, M$ imputed datasets which simulate features of the posterior distribution of $\hat{\theta}$. In addition, let $\bar{\theta}$ and $\mathcal{B}$ denote the posterior mean of $\hat{\theta}$ and its associated variance over multiple imputations, respectively. Then, for very large number of imputations M, the combined estimate gives the MI estimator $\bar{\theta}$ which is the posterior mean of $\hat{\theta}$,

$$\bar{\theta} = \frac{1}{M} \sum_{j=1}^{M} \hat{\theta}_j \approx \mathrm{E}(\hat{\theta} \mid y, r, z).$$

The total variability associated with this estimate is the sum of two components: the mean of the posterior variance $\mathcal{W}$ over multiple imputations and the variance of the posterior mean $\bar{\theta}$ over multiple imputations. The first component is obtained by the within-imputation variance

$$\overline{\mathcal{W}} = \frac{1}{M} \sum_{j=1}^{M} \mathcal{W}_j,$$

and the second component is obtained by the between-imputation variance

$$\mathcal{B} = \frac{1}{M-1} \sum_{j=1}^{M} (\hat{\theta}_j - \bar{\theta})(\hat{\theta}_j - \bar{\theta})^{\intercal} \approx \mathrm{Var}(\bar{\theta} \mid y, r, z),$$

hence, the poeterior variance of θ in (2.2) can be obtained by $\overline{\mathcal{W}} + \mathcal{B}$. Thus, the trio statistic $(\bar{\theta}, \overline{\mathcal{W}}, \mathcal{B})$ from the multiple imputed datasets provides the information needed to estimate the pair $(\hat{\theta}, \mathcal{W})$, and so to estimate θ (Van Buuren, 2012).

2.2.3 Theoretical properties

Proper MI. The goal of MI is to make a valid inference about θ. The Bayesian framework is useful for creating a MI procedure, however, multiple imputations are derived based on the assumption that models are correctly specified. That is, the validity of MI depends on how the imputations are created. For example, if multiple imputations are created arbitrarily, one is not likely to obtain valid inferences (Schafer, 1999). The imputations should, on average, give reasonable predictions for the missing data, and the associated variability between them should properly reflect the amount of uncertainty we have about them.

Recall that inferences based on MI are derived from the Bayesian framework treating θ and $\hat{\theta}$ as unobserved random variables that asymptotically follow the normal distribution given the observed values $\hat{\theta}_1, \ldots, \hat{\theta}_j$ and $\mathcal{W}_1, \ldots, \mathcal{W}_j$. Thus, for an infinitely large M, inference based on MI is valid if the inference from the complete data is valid. The second requirement for MI inference to be valid is that the MI procedure is proper (Rubin, 1987).

MI is proper if

1. $\bar{\theta}$ is approximately unbiased for $\hat{\theta}$ averaged over missingness mechanism: $\mathrm{E}(\bar{\theta} \mid y, z, r) = \hat{\theta}$.

2. The between-imputation variance is approximately unbiased: $\mathrm{E}(\mathcal{B} \mid y, z, r) = \mathrm{Var}(\bar{\theta} \mid y, z, r)$.

3. The within-imputation variance ($\overline{\mathcal{W}}$) is approximately unbiased averaged over the missingness mechanism: $\mathrm{E}(\overline{\mathcal{W}} \mid y, z, r) = \mathcal{W}$.

This means that, for infinitely large M, the trio statistic $(\bar{\theta}, \overline{\mathcal{W}}, \mathcal{B})$ of the incomplete data must provide valid inference for the pair statistic $(\hat{\theta}, \mathcal{W})$ of the observed data.

Therefore, for a multiple imputation method to belong to the class of *proper* MI, multiple imputations must yield a consistent asymptotically normal estimator of θ and an unbiased estimator of its asymptotic variance when based on Rubin's rule. By using heuristic arguments and several examples, Rubin (1987) concluded

that "*Conclusion 4.1. If imputations are drawn to approximate repetitions from a Bayesian posterior distribution of missing data under the posited response* (missingness) *mechanism and an appropriate model for the data, then in large samples the imputation method is proper*", where the posterior distribution of missing data is defined as

$$p(z \mid y, r) = \int p(z \mid y, r, \theta) p(\theta \mid y, r) \mathrm{d}\theta. \tag{2.3}$$

In other words, if the multiple imputations are created to approximate the observed posterior distribution of θ, then the advantageous properties of MI can be guaranteed (Van Buuren, 2012, p. 47) and the complete data inferences can be properly combined according to Rubin's rule. In summary, non-Bayesian MI (where its $\hat{\theta}_j$s are non-random draws from their observed posterior distribution) is an *improper* MI and its inference may not be based on Rubin's rule.

Choosing M. In practice, it is not feasible to produce infinite number of imputations. Hence, when M is finite, $\bar{\theta}$ and its variance estimate are subject to simulation error. Rubin (1987) recommended to account for this simulation error by adding a third component ($\mathcal{B}/M$) to the total variance estimate in Rubin's rule:

$$\overline{\mathcal{W}} + \left(1 + \frac{1}{M}\right)\mathcal{B}.$$

The choice of M could be interpreted as a trade-off between statistical efficiency versus computational efficiency. Clearly, the larger M is, the smaller the effect of simulation error to the total variance and the lower the computational efficiency.

Rubin (1987) suggested to choose M based on the fraction of information that is missing about θ due to missingness in order to achieve a reasonably small efficiency loss. Denote by γ the fraction of missing information matrix due to missingness. Based on the missing information principle (Orchard and Woodbury, 1972), the information matrix, $I(\theta)$, may be written as $I(\theta) = I_c(\theta) - \mathrm{E}_\theta[I_{z|y}(\theta)]$ where $I_c(\theta)$ is the complete data information and $I_{z|y}(\theta)$ is the missing data information,

and γ may be written as

$$\gamma = \mathrm{E}_\theta[I_{z|y}(\theta)]I_c^{-1}(\theta). \tag{2.4}$$

When θ contains a single parameter then γ is a scalar. Since MI inference is based on finite-M, this will result in a factor of γ/M efficiency loss (Rubin, 1987, p. 114). In the MI literature, γ can be estimated by $\hat{\gamma} = (1 + 1/M)\mathcal{B}/(\overline{\mathcal{W}} + (1 + 1/M)\mathcal{B})$ (Little and Rubin, 2002, p. 257). Also, Harel (2007) proposed to use the fraction of missing information for choosing M and derived asymptotic distribution of $\hat{\gamma}$ under MAR assumption. For scalar θ, this reduces to

$$\sqrt{M}(\hat{\gamma} - \gamma) \sim N\big(0, 2\gamma^2(1-\gamma)^2\big).$$

They proposed that $\hat{\gamma}$'s asymptotic distribution can be used to help choosing the number of imputations required to achieve reliable estimates for the fraction of missing information.

However, Graham *et al.* (2007) showed that the fraction of missing information cannot be reliably estimated unless M is sufficiently large, and furthermore, statistical power is more influenced by M rather than efficiency. They provided a practical guide for choosing M where they focused on the impact of different values of M on statistical power, especially in complicated situations where high statistical power is required, such as, for detecting a very small effect size (< 0.1). Furthermore, they concluded that for a missing proportion as high as 70%, only $M = 100$ imputations are required in order to obtain a power within 1% of the theoretical power. These results were used as a guide for selecting M in our applications in later chapters.

2.3 MLE via EM algorithm

The EM algorithm (Dempster *et al.*, 1977) was designed to find MLE estimates of parameters of a parametric model in an iterative manner when the observed data are incomplete. This procedure makes use of Fisher's identity, where the maximization of the unknown observed log-likelihood is replaced with the maxi-

mization of the conditional expectation of an associated complete log-likelihood:

$$\frac{\partial l(\theta; y, r)}{\partial \theta} = \mathrm{E}_\theta \left\{ \frac{\partial l(\theta; y, z, r)}{\partial \theta} \,\middle|\, y, r \right\}.$$

An EM iteration $\theta^{(t)} \to \theta^{(t+1)}$ consists of two steps. The E-step computes the expectation of conditional complete-data log-likelihood given the observed data (with respect to the imputation model at the current estimate of parameters),

$$Q(\theta \mid \theta^{(t)}) = \int \log \{p(y, z, r \mid \theta)\}\, p(z \mid y, r, \theta^{(t)}) \mathrm{d}z.$$

The M-step updates the estimates of parameters by maximization of the expectation function computed in the E-step,

$$\theta^{(t+1)} = \arg\max_{\theta \in \Theta} \; Q(\theta \mid \theta^{(t)}).$$

2.3.1 Stochastic EM algorithm

In situations where $Q(\theta \mid \theta^{(t)})$ is either analytically intractable (McCullagh, 1998) or computationally intensive (Ng and McLachlan, 2003), it is possible to replace analytical computation of $Q(\theta \mid \theta^{(t)})$ by a suitable approximation to this function, commonly via simulation methods such as MCMC. In this paradigm, stochastic versions of the EM algorithm were designed to numerically compute $Q(\theta \mid \theta^{(t)})$ by Monte Carlo approximation. As such, the E-step in the EM algorithm simplifies to the computation of the imputation model, $p(z \mid y, r, \theta^{(t)})$, and simulation of the missing data $z^{(t)}$. In other words, the E-step turns into an imputation step (I-step) where multiple imputations are drawn from

$$z^{(j)} \sim p(z \mid y, r, \theta^{(t)}), \quad j = 1, \ldots, M,$$

given a current approximation to MLE ($\theta^{(t)}$), to approximate $Q(\theta \mid \theta^{(t)})$ as a Monte Carlo average,

$$Q(\theta \mid \theta^{(t)}) \simeq \frac{1}{M} \sum_{j=1}^{M} \log p(y, r, z^{(j)} \mid \theta).$$

A popular stochastic version of EM is the Monte Carlo EM (MCEM, Wei and Tanner, 1990) where $M \geq 2$ is fixed throughout the iterations of the I-step and M-step. The MCEM algorithm, under mild conditions, converges to the MLE, and inference about θ may be made based on the final M imputations. It has also been suggested, instead of a fixed value for M, to start the MCEM algorithm with small number of imputations and increase M with the number of iterations. This approach will ensure higher computational efficiency in cases where the initial estimates of θ might be far from the true value.

A special case of the EM algorithm is the Stochastic EM (StEM, Broniatowski *et al.*, 1983; Celeux and Diebolt, 1985, 1987), summarized below. StEM, without aiming to produce any approximate computation of $Q(\theta \mid \theta^{(t)})$, imputes the missing data *only once* in the I-step until the algorithm converges to its stationary distribution, whose mean is close to the MLE (Diebolt and Ip, 1996). The random sequence of $\{\theta^{(t)}\}$ generated by the StEM does not converge pointwise to the MLE, but, under mild conditions, does converge in distribution (Diebolt and Celeux, 1993). After the algorithm converges, multiple imputations are generated by running extra $M \geq 2$ iterations to sample from the stationary distribution, and the sample mean gives an approximate MLE of the observed likelihood.

StEM algorithm

0. Fix $\theta^{(t)}$ in Θ, $t = 0$
1. Draw $z^{(t+1)}$ from $p(z \mid y, r, \theta^{(t)})$
2. $\theta^{(t+1)} = \arg\max\ p(y, r, z^{(t+1)} \mid \theta)$
3. Set $t = t + 1$ and repeat steps 1–2 until convergence at $t = T$
4. Repeat steps 1–2 for extra $M \geq 2$ iterations, $t = T + 1, \ldots, T + M$, to create multiple imputations
5. Combine the results of the final M iterations by Louis' method

The StEM estimator has been shown to be an asymptotically normal, unbiased, and consistent estimator of θ when considering models from the exponential family (Diebolt and Celeux, 1993) and in mixture models on the basis of numerical experiments (Celeux *et al.*, 1996). Asymptotic properties of the StEM estimator are given in Wang and Robins (1998) and in Nielsen (2000).

2.3.2 Combination rules

StEM's combination rules are carried out as follows. At each iteration, multiple imputations are randomly drawn from the imputation model given the current MLE of θ. Let $\hat{\theta}_1, \ldots, \hat{\theta}_j$ and $\mathcal{W}_1, \ldots, \mathcal{W}_j$, $j = 1, \ldots, M$, be the MLEs of θ and their variances, respectively, obtained from the next M iterations after the StEM algorithm converges. These final M imputations are implemented as multiple imputations in combination rules as shown in Diebolt and Ip (1996) in a manner discussed below.

Let $\bar{\theta} = M^{-1} \sum_{j=1}^{M} \hat{\theta}_j$ be the StEM estimator which is the average of $\hat{\theta}_j$s over multiple imputed datasets. Wang and Robins (1998) and Nielsen (2000) showed that, for a sufficiently large M, $(\bar{\theta} - \theta) \sim N(0, I^{-1}(\theta))$ where $I(\theta)$ denotes the Fisher information matrix. As such, the variance of the StEM estimator may be obtained based on the Louis' method (Louis, 1982),

$$I(\theta) = \mathrm{E}\left[\frac{-\partial^2}{\partial\theta\partial\theta^\top} \log p(y, z, r \mid \theta)\middle| y, r\right] - \mathrm{Var}\left[\frac{\partial}{\partial\theta} \log p(y, z, r \mid \theta)\middle| y, r\right], \tag{2.5}$$

and by replacing $\mathrm{E}[. \mid y, r]$ and $\mathrm{Var}[. \mid y, r]$ with their bootstrap estimates: from the difference between the complete information matrices of the $\hat{\theta}_j$s averaged over multiple imputed datasets *and* the variance of their respected score functions between multiple imputed datasets. von Hippel (2012) showed that the variance of $\bar{\theta}$ based on the inverse of $I(\theta)$ in Equation (2.5) can be simplified into

$$I^{-1}(\theta) = \overline{\mathcal{W}}^\top (\overline{\mathcal{W}} - \mathcal{B})^{-1} \overline{\mathcal{W}}$$

where $\overline{\mathcal{W}}$ and $\mathcal{B}$ denote the within and between-imputation variances of $\hat{\theta}_j$s, respectively.

2.3.3 Theoretical properties

Stochastic versions of EM are motivated by overcoming the limitations of the EM algorithm. The EM algorithm strongly depends on its starting value, furthermore, it can have a very slow convergence rate and can converge to a saddle point of

the loglikelihood function rather than a local maxima (McLachlan *et al.*, 2004; Ng *et al.*, 2012). Stochastic versions of EM, on the other hand, can avoid these limitations due to their underlying stochasticity (Celeux *et al.*, 1996). In addition to these advantages, the StEM algorithm is computationally more efficient since it is motivated by simulation of complete loglikelihoods instead of providing a Monte Carlo approximation of $Q(\theta \mid \theta^{(t)})$. However, the StEM algorithm loses some efficiency for small values of M due to its maximize-then-average strategy.

Choosing M. Nielsen (2000) studied asymptotic behaviour of stochastic versions of EM and showed that for finite M, the asymptotic variance of their estimator, $\bar{\theta}$, is

$$(\bar{\theta} - \theta) \sim N\Big(0, I^{-1}(\theta) + \frac{1}{M} I^{-1}(\theta)(I - \{I + \gamma\}^{-1}) + \eta\Big) \tag{2.6}$$

where γ is the fraction of missing information as defined in (2.4), $\eta = 0$ for the MCEM estimator, and for the StEM estimator we have

$$\begin{aligned}\eta = &\frac{2}{M} I^{-1}(\theta)\left(I - \{I + \gamma\}^{-1}\right)\gamma(I - \gamma)^{-1} \\ &- \frac{2}{M^2} I^{-1}(\theta)\left(I - \{I + \gamma\}^{-1}\right)\gamma(I - \gamma)^{M}(I - \gamma)^{-2}.\end{aligned}$$

Again, it is not straightforward how to choose M, as it will require a reliable estimate of γ to obtain an acceptable efficiency loss. Meng and Rubin (1991) proposed an approach to estimate γ in the context of assessing the convergence rate of the EM algorithm. To estimate γ they proposed to run the EM algorithm after convergence (supplementary EM) and compute each ij^{th} element of γ in turn by keeping constant all the other elements at the MLE and using the following ratio

$$\hat{\gamma}_{ij} = \frac{\theta_j^{(t+1)} - \hat{\theta}_i}{\theta_j^{(t)} - \hat{\theta}_i}.$$

Nielsen (2000) proposed to estimate the largest eigenvalue of γ for choosing M. For example, they showed that, based on their result in (2.6), the asymptotic

relative efficiency of the MCEM estimator is bounded by

$$\big[1 - 1/(1+\lambda)\big]/M \leq 1/(2M)$$

where λ is the largest eigenvalue of γ. Let λ^* be an upper bound on λ. By specifying how small the efficiency loss that we are willing to accept, say δ, should be, we may choose M by

$$M \geq 1/\big\{\delta\big[1 - 1/(1+\lambda^*)\big]\big\}, \tag{2.7}$$

which ensures that the loss in efficiency is bounded by δ.

There are a few studies in the literature that have discussed the estimation of λ (*e.g.,* in Fraley (1991)), however, Nielsen (2000) suggested that it would be possible to use a looser bound for choosing M. Based on the closer bound in (2.7), a loose bound for M could be $M \geq 1/(2\delta)$ in the MCEM algorithm, and analogously derived from (2.6), $M \geq 2/(3\delta)$ in the StEM algorithm (Nielsen, 2000, proposition 4).

So far, estimation of the fraction of missing information (either using the largest eigenvalue or the above matrix) seems to be the most important criteria in choosing M in both MI and MLE literatures. However, this estimation itself is dependent on M based on both Rubin's rule or the missing information principle, where MI and MLE offer two different estimators for γ. Throughout this thesis, we combined Graham *et al.* (2007)'s heuristic guidelines form the MI literature with Nielsen (2000)'s loose bound approach from the MLE literature and set $M = 100$ for our simulation studies and examples in later chapters. This value was selected since we have missing proportions of $< 70\%$ everywhere, which results in only 0.0067 asymptotic efficiency loss following Nielsen (2000)'s loose bound approach.

2.4 Equivalence of MI and StEM

Studying Box 1 below, it is evident that the MI and StEM algorithms described above are almost identical – both iterate between imputation from the current

Box 1. MI (proper) and StEM (improper) algorithms, note that they only differ at step 2.

MI (“*proper*”):	**StEM (“*improper*”)**:
0. Fix $\theta^{(0)}$ in Θ	0. Fix $\theta^{(0)}$ in Θ
1. $z^{(t+1)} \sim p(z \mid y, r, \theta^{(t)})$	1. $z^{(t+1)} \sim p(z \mid y, r, \theta^{(t)})$
2. $\theta^{(t+1)} \sim p(\theta \mid y, r, z^{(t+1)})$	2. $\theta^{(t+1)} = \arg\max\ p(y, r, z^{(t+1)} \mid \theta)$
3. Repeat steps 1–2 until convergence (at $t = T$)	3. Repeat steps 1–2 until convergence (at $t = T$)
4. $\bar{\theta} = \frac{1}{M} \sum_{j=T+1}^{T+M} \theta^{(j)}$	4. $\bar{\theta} = \frac{1}{M} \sum_{j=T+1}^{T+M} \theta^{(j)}$

model (Step 1: “I step”) and updating the imputation model (Step 2: “P step”), then following convergence at iteration $t = T$. Finally, both algorithms involve averaging estimates obtained from a set of M ensuing iterations as the finial point estimate of θ:

$$\bar{\theta} = \frac{1}{M} \sum_{j=T+1}^{T+M} \theta^{(j)}.$$

The only difference between the two algorithms is in how θ is updated at Step 2 – by using random draws from the current posterior, or by using the maximizer of the current estimate of the likelihood function. Thus, reviewing the underlying drive for this difference could be helpful to understand how similar these two algorithms can actually be.

MI views θ as random (with a prior distribution) and samples values from the observed posterior. These samples, which are obtained from the posterior are used to impute the missing values z, and are eventually averaged to approximate the mean of the posterior distribution of θ. StEM views θ as fixed, and at each step uses an estimate of it in the model that imputes missing values z, and eventually averages these for a final estimate of θ. From the MI point of view, this can be understood as approximating the *mode* of the posterior distribution with flat priors, rather than approximating the *mean*. However, both algorithms assume their estimators are asymptotically normal (Rubin, 1987; Diebolt and Celeux, 1993), which would imply that the mean and mode would converge asymptotically.

The combination of Step 1 and 2 in both algorithms serves the purpose of creating

multiple imputations by randomly drawing from the conditional predictive distribution of missing data $p(z \mid y, r)$. Since $p(z \mid y, r)$ is usually unknown, multiple imputations are drawn from $p(z \mid y, r, \theta^{(t)})$, where $\theta^{(t)}$ is the current approximation to θ. MI uses random draws from the current posterior to obtain a current approximation to $p(z \mid y, r)$ whereas StEM draws from $p(z \mid y, r, \theta^{(t)})$ where $\theta^{(t)}$ is the current approximation to the maximizer of the likelihood function. While the former is a proper MI (Rubin, 1987), the latter is not since Equation (2.3), and therefore Equation (2.2), will no longer be satisfied.

While Rubin's combination rule (Rubin, 1987) enables calculation of approximate standard errors when using proper MI, these are not available in the improper case. StEM nevertheless comes with standard approaches to estimate standard errors via Louis' method, which interestingly, has a similar form as Rubin's rule (see Sections 2.2.2 and 2.3.2). Hence, if we create multiple imputations by randomly drawing from the conditional predictive distribution of the missing data $p(z \mid y, r, \theta^{(t)})$, $\theta^{(t)}$ being the approximation to current MLE of θ, StEM and improper MI are equivalent.

2.5 Gains from the equivalence

The equivalence between StEM and improper MI allows standard likelihood machinery to be used to improve MI's performance. This connection changes our view of MI and provides the possibility of accessing a range of likelihood-based tools in situations where MI's performance could be improved in a likelihood-based framework.

One important gain, for example, is when a maximum likelihood framework can provide the analyst with model selection tools for choosing the best imputation model, and can enable further insights into the consequences of imputation model misspecification. In the MI literature, there are no diagnostic criteria for how to choose between imputation models. In a heuristic manner, it is recommended to either specify a multivariate normal distribution as a joint model for missing data (Schafer, 1997) or specify a univariate conditional imputation model for each

missing variable (Van Buuren, 2012). Also, it is recommended to add as many variables in the imputation model as possible, with at least as many variables as presented in the substantive analysis model of interest (Rubin, 1996; Collins *et al.*, 2001). Furthermore, it is not clear how to choose between different missing data mechanisms except for performing a sensitivity analysis to explore the impact of the different assumptions for missingness mechanism (Carpenter *et al.*, 2007; Sterne *et al.*, 2009). In Chapter 3 we show that available and well-known information-based criteria in the maximum likelihood literature, which enjoy good statistical properties, can be used to select an imputation model. We investigate these tools using simulation studies and apply it to several real-data examples.

Another gain from the equivalence is in hypothesis testing and in goodness-of-fit where likelihood ratio tests (LRT) are commonly used. Under a maximum likelihood framework, these inferential tools could then be easily employed when missing data are present. In the MI literature, hypothesis testing based on multiple imputed datasets was proposed to obtain a modified Wald test statistic (Rubin, 1987). Subsequent work by Meng and Rubin (1992) developed a pooling procedure for a likelihood ratio test with MI for nested models, using the asymptotic relationship between the Wald test and the likelihood ratio test statistics. Although this approach works well, it can be quite cumbersome to implement in practice. Using connections with maximum likelihood, a likelihood ratio statistic could be directly constructed for MI. In Chapter 4, by means of simulation studies, we show that this likelihood ratio statistic has improved performance over other statistics in the MI literature.

Furthermore, in the presence of both missingness and measurement error in explanatory variables, the equivalence allows us to combine MI with functional measurement error models to account for both missingness and measurement errors in the inference. The combined presence of missingness and measurement error in explanatory variables may have a double effect on statistical analyses if not dealt with, since both missingness and measurement error may cause bias in estimates of regression coefficients and loss of power if not dealt with. In the MI literature, there are few studies on how to deal with this double effect. The likelihood-based

framework for dealing with both missingness and measurement error is very attractive. For example, Blackwell *et al.* (2017a,b) applied EM-type algorithms to account for both missingness and measurement errors where they treated variables subject to measurement error as missing while incorporating available information about their error contaminations in the model. Specifically, Blackwell *et al.* (2017a) used EM with bootstrapping as a multiple imputation method. However, they used Rubin's combination rules in inferences. These studies were restricted to the assumption of missingness/measurement mechanism ignorability. In Chapter 5 we show that the equivalence between StEM and improper MI allows us to combine functional measurement error modelling methods such as Simulation extrapolation and Corrected score with MI to deal with the combined effect of missingness with measurement error in explanatory variables in a simpler manner. Again, we investigate this result using simulation studies and real-data examples.

It is worth noting that the gains from this connection are not limited to the applications studied in this thesis. There are other potential areas where ideas from the maximum likelihood literature can be taken across to MI. For example, suppose that we would like to predict myocardial infarction in patients with observed blood pressure, body mass index, age and gender, but with missing cholesterol levels. In the MI literature, there are no guidelines on how to approach prediction in the presence of missing data, and most of the methods suggested are ad hoc (Wood *et al.*, 2015). However, there is a clear guidance on how to carry out prediction in a likelihood based framework and the close connection between MI and StEM could provide clarity in this field. Also, as mentioned in Section 2.2, currently there are only heuristic approaches for choosing M in the MI literature. However, StEM's asymptotic results studied by Nielsen (2000) and Wang and Robins (1998) could provide complementary tools on this topic.

Chapter 3

Imputation model selection with missing data

3.1 Background

The literature on imputation model selection within the MI framework is surprisingly sparse given its potential application range. When using MI, careful consideration of which imputation model to select is needed, because using the wrong imputation model can result in incorrect inferences and misleading conclusions (Fay, 1991, 1996; Schomaker and Heumann, 2014). There are several guidelines for the specification of the imputation model but they tend to be ad hoc and based on heuristic arguments (Schafer, 1997; Graham, 2009; Van Buuren, 2012) rather than providing an overall valid framework for imputation model selection.

Rubin (1987) suggested an analysis to assess the sensitivity of MI inference to alternative models for missingness. For this procedure, imputations are created under different assumptions of missingness and a sensitivity analysis is carried out to see how inferences may vary (Van Buuren, 2012). Sensitivity of MI results are often assessed using a weighting approach (Carpenter *et al.*, 2007) or the pattern-mixture approach (Little, 1993; Little and Rubin, 2002). Recently, Andridge and Thompson (2015) proposed a variable selection procedure to select the best imputation model using the fraction of missing information as the selection criterion. The fraction of missing information of the imputation model is estimated from

an appropriate Proxy pattern-mixture model, as a function of estimates of first– and second order moments. The Proxy pattern-mixture technique reduces a set of variables to a single proxy variable, which is then used for imputation.

A key focus of this chapter is to provide a flexible, data-driven approach for choosing the imputation model. Specifically, we exploit the equivalence between MI and StEM to propose imputation model selection using standard information criteria such as Akaike's Information Criterion (AIC) and Bayesian Information Criterion (BIC), computed using the observed data likelihood. We develop a theorem which shows that BIC preserves its usual property of consistency in model selection when selecting an imputation model. We also illustrate this property via simulation.

In Section 3.2 we develop AIC and BIC for choosing the best imputation model and study their properties. We then apply the methods and present the results for three real data examples in health and ecological research studies (see Section 1.2.1 for further details). Two simulation studies are given in Section 3.3. Finally, we provide a proof for Theorem 1 in Section 3.8.

3.2 Information criteria

The connection between StEM and improper MI enables application of well-known information criteria such as AIC and BIC for choosing the best imputation model among a set of candidate imputation models. In the absence of missing data, AIC aims to find the best approximating model to the unknown correct model, whereas BIC aims to identify the correct model among a set of candidates (Acquah, 2010). Both AIC and BIC use the observed log-likelihood for some candidate model fit to assess its goodness-of-fit with the addition of a penalty term to account for model complexity. Let $p(y, r \mid \theta)$ be the observed likelihood and denote $|\theta|$ as the number of model parameters. We write $\mathrm{AIC} = -2\sup_\theta \log p(y, r \mid \theta) + 2|\theta|$, and $\mathrm{BIC} = -2\sup_\theta \log p(y, r \mid \theta) + \log(n) \times |\theta|$. Below, we will present a key property when applying BIC to imputation model selection.

In Section 3.8 we give a proof of the following theorem, which shows that BIC is consistent for imputation model selection, that is, it chooses the correct imputation

model with probability approaching 1 as the sample size n goes to ∞:

Theorem 1. *Suppose $\mathbb{M}_0$ is the imputation model chosen by BIC and $\mathbb{M}^p$ is a finite set of the most parsimonious correct models. If Assumptions A1–A4 (see Section 3.8) are satisfied, then*

$$\Pr(\mathbb{M}_0 \in \mathbb{M}^p) \to 1 \;\; as \;\; n \to \infty.$$

3.2.1 Why do information criteria work?

We can explain how the observed log-likelihood is informative about the imputation model using an approximation technique that, interestingly, is equivalent to that used in variational approximation of likelihood functions (Ormerod and Wand, 2010). Variational approximation is a (trade-off) method for optimizing a log-likelihood function while enhancing its tractability, hence, making approximate inference for parameters of a model (see Section 1.5.2). In the missing data context, an incorrect imputation model can be understood as a variational approximation to the observed log-likelihood. The approximations are indicated by the Kullback–Leibler divergence (Kullback and Leibler, 1951) of the specified imputation model from the true imputation model.

Suppose $q(z \mid y, r, \theta)$ is a specified imputation model and $p(z \mid y, r, \theta)$ is the true imputation model. Misspecification of the imputation model implies that $q(z \mid y, r, \theta)$ diverges from $p(z \mid y, r, \theta)$. The divergence of $q(z \mid y, r, \theta)$ from $p(z \mid y, r, \theta)$ is assessed by the Kullback–Leibler divergence, denoted by $KL(q||p)$, and is always nonnegative. A zero value would only occur when the imputation model is not misspecified, whereas a positive value would imply that the imputation model is misspecified, and the further $q(z \mid y, r, \theta)$ is from $p(z \mid y, r, \theta)$, the larger this divergence is, *i.e.,* the greater is $KL(q||p)$.

Following a similar approach as given in Ormerod and Wand (2010), we have

$$
\begin{aligned}
\log p(y,r \mid \theta) &= \log p(y,r \mid \theta) \int q(z \mid y,r,\theta)\mathrm{d}z \\
&= \int q(z \mid y,r,\theta) \log p(y,r \mid \theta)\mathrm{d}z \\
&= \int q(z \mid y,r,\theta) \log \left(\frac{p(y,z,r \mid \theta)}{q(z \mid y,r,\theta)} \frac{q(z \mid y,r,\theta)}{p(z \mid y,r,\theta)} \right) \mathrm{d}z \\
&= \int q(z \mid y,r,\theta) \log \left(\frac{p(y,z,r \mid \theta)}{q(z \mid y,r,\theta)} \right) \mathrm{d}z - \int q(z \mid y,r,\theta) \log \left(\frac{p(z \mid y,r,\theta)}{q(z \mid y,r,\theta)} \right) \mathrm{d}z \\
&= \int q(z \mid y,r,\theta) \log \left(\frac{p(y,z,r \mid \theta)}{q(z \mid y,r,\theta)} \right) \mathrm{d}z + \int q(z \mid y,r,\theta) \log \left(\frac{q(z \mid y,r,\theta)}{p(z \mid y,r,\theta)} \right) \mathrm{d}z \\
&= \int q(z \mid y,r,\theta) \log \left(\frac{p(y,z,r \mid \theta)}{q(z \mid y,r,\theta)} \right) \mathrm{d}z + KL(q\|p) \\
&\geq \int q(z \mid y,r,\theta) \log \left(\frac{p(y,z,r \mid \theta)}{q(z \mid y,r,\theta)} \right) \mathrm{d}z. \qquad (3.1)
\end{aligned}
$$

The maximum value of the lower bound in (3.1) over q is obtained when $q(z \mid y,r,\theta) = p(z \mid y,r,\theta)$ (that is, when the Kullback–Leibler divergence of q from p is at a minimum, or $KL(q\|p) = 0$), since

$$
\int q(z \mid y,r,\theta) \log \left(\frac{p(y,z,r \mid \theta)}{q(z \mid y,r,\theta)} \right) \mathrm{d}z = \int p(z \mid y,r,\theta) \log \left(\frac{p(y,z,r \mid \theta)}{p(z \mid y,r,\theta)} \right) \mathrm{d}z = \log p(y,r \mid \theta).
$$

Therefore, the better the specified imputation model, the lower $KL(q\|p)$, and hence, the higher $\int q(z \mid y,r,\theta) \log \left(\frac{p(y,z,r|\theta)}{q(z|y,r,\theta)} \right) \mathrm{d}z$. As such, information criteria AIC and BIC based on observed log-likelihoods would be able to reflect the goodness-of-fit of the chosen imputation model.

3.3 Simulation study

We present two simulation studies to investigate the performance of AIC and BIC as imputation model selection criteria where we have: (I) univariate missing variable; and (II) multivariate missing variables.

3.3.1 Univariate missing variable

We are interested in a linear regression model, $Y_i = \beta_0 + \beta_1 X_{1i} + \epsilon_i$, where Y_i, $i = 1, \ldots, n$, is the response variable and the errors are $\epsilon_i \overset{i.i.d.}{\sim} N(0, 1)$. The predictor X_{1i} is partially observed, and suppose that X_{2i} and X_{3i} are two fully observed auxiliary variables where $(\log X_{1i}, \log X_{2i}, \log X_{3i}) \sim N_3(\mu, \Sigma)$ with a mean vector $\mu = (0, 0, 0)^\intercal$ and a covariance matrix

$$\Sigma = \begin{pmatrix} 1 & 0.5 & 0 \\ 0.5 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

In this scenario, X_{1i} is correlated with X_{2i} with a moderate correlation of 0.5, but is independent of X_{3i}. Furthermore, Y_i is conditionally independent of X_{2i} given X_{1i}.

We imposed 30% left-censoring in X_1 where the limit of detection (L) is treated as a fixed number and set equal to the corresponding quantile. Observations below this threshold were removed by setting them to NA. Because the limit of detection is fixed and not stochastic, this type of censoring belongs to the class of Type-I censoring which Heitjan and Rubin (1991) showed to be MAR (more specifically, coarsened at random), and thus, the missingness mechanism is ignorable. In other words, although we use a truncated lognormal distribution to draw imputations for missing values, the missingness mechanism can be ignored in the imputation model as the missingness within the truncated threshold does not depend on X_1 once we have a reasonable estimate of the limit of detection. See Heitjan and Rubin (1991) for more details.

We ran 500 simulations and set $\beta = (1, 1)^\intercal$. We used the squared distance of parameters between the $(t+1)^{th}$ and the t^{th} iterations as a convergence criterion for the StEM algorithm, and set the convergence threshold to 10^{-4}. Furthermore, the number of multiple imputations was set to $M = 100$. To investigate the performance of information criteria for imputation model selection, the missing values $X_{mis,1}$ were imputed under the following three candidate models:

Model 1: $p(X_{mis,1i} \mid X_{mis,1i} \leq L, X_{obs,1i}, Y_i, X_{2i}, X_{3i})$,

Model 2: $p(X_{mis,1i} \mid X_{mis,1i} \leq L, X_{obs,1i}, Y_i, X_{2i})$,

Model 3: $p(X_{mis,1i} \mid X_{mis,1i} \leq L, X_{obs,1i}, Y_i, X_{3i})$,

Model 4: $(X_{mis,1i} \mid X_{mis,1i} \leq L, X_{obs,1i}, Y_i, X_{2i})$ is truncated normal.

Results in Table 3.1 show the proportion of times the corresponding model is chosen based on each information criterion for various sample sizes of $n = \{50, 100, 1000\}$. These results indicate that for even a small sample size of 50, both AIC and BIC are able to choose the correct model (Model 2) at least 89.6% of the time, with misspecified models (Model 3 and Model 4) selected rarely, and selected at a rate that went to zero as sample size increased. The over-fitted model (Model 1) was chosen by AIC 7-15% of the time, increasing with sample size. For BIC, the rate at which the over-fitted model was chosen by BIC went to zero as sample size increased, as expected from Theorem 1. These results align with classical results for complete data cases, where several studies have shown that AIC overfits (asymptotically) (Shibata, 1976; Bozdogan, 1987; Hurvich and Tsai, 1989) while in contrast, BIC can be consistent for model selection (Schwarz, 1978; Nishii, 1984).

Table 3.1. Proportion of times (%) the information criterion chooses the fitted imputation model for different sample sizes in 500 simulated datasets. Model 2 (in bold) is the correct model, which was selected most of the time in each simulation. Note that the proportion of times this model was chosen by BIC went to one as sample size increased, as expected under Theorem 1.

		Model 1	**Model 2**	Model 3	Model 4
n=50	AIC	7.6	**89.6**	2.8	0.0
	BIC	1.0	**95.4**	3.6	0.0
n=100	AIC	8.8	**91.2**	0.0	0.0
	BIC	0.0	**99.8**	0.2	0.0
n=1000	AIC	15.6	**84.4**	0.0	0.0
	BIC	0.0	**100**	0.0	0.0

3.3.2 Multivariate missing variable

Next we investigated the case where missingness was of a higher dimension, and the response was non-normal. Following Example 4.1.2 from Ibrahim *et al.* (1999) we now suppose that the response variables Y_i, $i = 1, \ldots, n$, are independent fully

observed Bernoulli variables and we are interested in the analysis of a logistic regression model $\mathrm{E}(Y_i \mid X_{1i}, X_{2i}, \beta) = \exp(X_i\beta)/\{1+\exp(X_i\beta)\}$ where $\beta = (\beta_0, \beta_1, \beta_2)^\top$ and $X_i = [1\ X_{1i}\ X_{2i}]$. Suppose that the predictors, $X_i = (X_{1i}, X_{2i})$, $i = 1, \ldots, n$, are partially observed variables from a bivariate normal distribution $N_2(\mu, \Sigma)$ with

$$\mu = \begin{pmatrix} 0.5 \\ -0.5 \end{pmatrix} \quad \text{and} \quad \Sigma = \begin{pmatrix} 0.25 & 0.125 \\ 0.125 & 0.25 \end{pmatrix}.$$

Also, suppose that $r_i = (r_{1i}, r_{2i})$ is the missingness indicator vector of (X_{1i}, X_{2i}) – *e.g.*, $r_i = (0, 0)$ if X_{1i} and X_{2i} are both missing and $r_i = (0, 1)$ if X_{1i} is missing and X_{2i} is observed. Let ϕ_1 and ϕ_2 be the probabilities of $r_{1i} = 1$ and of $r_{2i} = 1$, respectively, which satisfy

$$\mathrm{logit}(\phi_1) = \log(\phi_1)/\log(1-\phi_1) = \phi_{10} + \phi_{11}X_{1i} + \phi_{12}X_{2i} + \phi_{13}Y_i$$

and

$$\mathrm{logit}(\phi_2) = \phi_{20} + \phi_{21}X_{1i} + \phi_{22}X_{2i} + \phi_{23}Y_i + \phi_{23}r_{1i}.$$

We are interested in the comparison of the following two imputation models:

Model 1: $p(X_{mis,i} \mid X_{obs,i}, Y_i, r_i)$ and

Model 2: $p(X_{mis,i} \mid X_{obs,i}, Y_i)$,

where Model 1 assumes that the data are MNAR and missingness is nonignorable whereas Model 2 assumes that data are MAR and ignores the missingness.

We ran 500 simulations with $n = 250$, $\beta = (1, 1, -1)^\top$, $\phi_1 = (1, -1, 1, 1)^\top$ and $\phi_2 = (1, -1, 1, 1, -0.5)^\top$ where on average we obtained about 19% missing data in X_1, 25% in X_2 and 6% in both. Also, to draw imputations at the ith iteration of StEM, we used Metropolis-within-Gibbs sampler (Tierney, 1999), see also Section 1.5.4, to generate a sample from

$$p(X_{mis,i} \mid X_{obs,i}, Y_i, r_i, \gamma_1^{(t)}) \propto p(r_i \mid Y_i, X_{mis,i}, X_{obs,i}, \phi^{(t)})p(Y_i \mid X_{mis,i}, X_{obs,i}, \beta^{(t)})p(X_{mis,i}, X_{obs,i} \mid \alpha^{(t)}) \quad (3.2)$$

for Model 1, and a sample from

$$p(X_{mis,i} \mid x_{obs,i}, Y_i, \gamma_2^{(t)}) \propto p(Y_i \mid X_{mis,i}, X_{obs,i}, \beta^{(t)}) p(X_{mis,i}, X_{obs,i} \mid \alpha^{(t)})$$

for Model 2 where $\alpha = (\mu, \Sigma)$, $\gamma_1 = (\phi, \beta, \alpha)$ and $\gamma_2 = (\beta, \alpha)$.

Once again, we used the squared distance of the parameters between the $(t+1)^{th}$ and t^{th} iterations as the convergence criterion and set the convergence threshold to 10^{-4} with the number of multiple imputations set to $M = 100$.

Table 3.2 shows the proportion of times each model was chosen based on the corresponding information criterion. This table shows that AIC and BIC were able to choose the correct model (Model 1) 99.4% and 95.2% of the times, respectively. These results are consistent with the previous simulation results (*cf.* Section 3.3.1) as well as with our theoretical result (*cf.* Section 3.2), indicating that in situations where AIC and BIC are applicable, they perform satisfactorily for imputation model selection.

Table 3.2. Proportion of times (%) the information criterion chooses the corresponding model across 500 simulated datasets. Data were generated under Model 1, which assumes MNAR, whereas Model 2 assumed data were MAR. Note that both approaches were able to recover the correct model with high probability.

	Model 1	Model 2
AIC	**99.4**	0.6
BIC	**95.2**	4.8

Perhaps the importance of imputation model selection is better demonstrated by investigating its impact on the post-selection inference of data. Table 3.3 shows the sample average of estimates of β_j, denoted as $\bar{\beta}_j$, $j = 0, 1, 2$, for the corresponding imputation model averaged over 500 simulations *and* its mean square error, $\text{MSE}_j = (\bar{\beta}_j - \beta_j)^2 + s_j^2$, where s_j is the simulated standard error of $\bar{\beta}_j$. This table shows that more accurate estimates and smaller mean squared errors are obtained when using the correct imputation model (Model 1).

Table 3.3. $\bar{\beta}_j$ and mean square error (in parenthesis) for different imputation models across 500 simulations.

	$\beta_0 = 1$	$\beta_1 = 1$	$\beta_2 = -1$
Model 1	1.008 (0.027)	1.014 (0.083)	-1.018 (0.084)
Model 2	1.521 (0.342)	1.288 (0.177)	-1.295 (0.186)

3.4 Survival of infants data revisited

Assuming that these data are MAR and the missingness is ignorable, Little and Rubin (2002) applied the EM algorithm to fit various models, such as $\{SC, SP, PC\}$, $\{SP, SC\}$, and $\{SC, PC\}$ where, for example, $\{SC\}$ denotes a model with all the main effects of Survival, Prenatal and Clinic *and* the interaction effect between Survival and Clinic. The goodness-of-fit using likelihood-ratio tests was then assessed for each candidate model. Meng and Rubin (1992) applied Bayesian MI with a full (saturated) model where they tested the null models $\{SC, PC\}$ and $\{S, P, C\}$ (a main effects only model) against the full model using their proposed pooled likelihood-ratio test developed for MI. Both approaches concluded that Survival is related to Clinic, but conditional on Clinic, Survival and Prenatal care are independent indicating that $\{SC, PC\}$ is the best parsimonious fitted model.

This example can also be viewed as a problem of imputation model selection, where the response variable is the missing variable. Thus, the model that we choose to fit to the data can be used as the imputation model in the MI algorithm. As such, we will have imputation model candidates $\{SC, SP, PC\}$, $\{SP, SC\}$, $\{SC, PC\}$, and so on. We fitted five competing imputation models as follows:

Model 1: $\{SC, SP, PC\}$,

Model 2: $\{SC, PC\}$,

Model 3: $\{SP, PC\}$,

Model 4: $\{SP, SC\}$,

Model 5: $\{S, P, C\}$.

To demonstrate the model fitting procedure, the StEM/improper MI algorithm (see

Section 2.3.1) can be applied to this problem by the following iterative steps:

1. Estimate initial value $\hat{\pi}_{ijk}^{(0)}$ based on the observed counts.

2. For t^{th} iteration, $t = 1, \ldots, T$,

   - Simulate $\hat{n}_{ijk}^{(t)}$ from $Bin(n_{ij.}, \hat{\pi}_{ijk}^{(t-1)})$.

   - Re-estimate π_{ijk} as $\hat{\pi}_{ijk}^{(t)} = \arg\max\ \ell(\pi)$, where $\ell(\pi) = \sum_i \sum_j \sum_k n_{ijk} \log \pi_{ijk}$ is the complete log-likelihood of a multinomial model. For instance, under Model 2, imputations for the partially classified counts of the $ij1^{th}$ cell are drawn from $\hat{n}_{ij1} \sim Bin(n_{ij.}, \hat{\pi}_{ij1})$ where $\hat{\pi}_{ij1} = \hat{\mu}_{ij1}/(\hat{\mu}_{ij1} + \hat{\mu}_{ij2})$ and $\hat{n}_{ij2} = n_{ij.} - \hat{n}_{ij1}$.

3. Calculate the AIC/BIC for the fitted model using criteria presented in Section 3.2.

Table 3.4 shows the AIC and BIC for the above-mentioned imputation models based on $M = 100$ multiple imputations. Also, the convergence threshold for the algorithm was set to 10^{-4}. AIC and BIC both favoured Model 2, in line with the results of Little and Rubin (2002) and Meng and Rubin (1992).

Table 3.4. Information criteria for candidate imputation models in Survival of infants data. Both AIC and BIC favoured Model 2, $\{SC, PC\}$, in line with previous analyses (Little and Rubin, 2002; Meng and Rubin, 1992).

	Model 1	**Model 2**	Model 3	Model 4	Model 5
AIC	27.02	**25.12**	33.29	156.71	168.28
BIC	63.60	**57.13**	65.30	188.71	191.14

Table 3.5 shows the estimated cell probabilities for the above-mentioned imputation models. These estimates differ under each imputation model. For example, the percentage of infants dying with less prenatal care in clinic A, $\pi_{111} \times 100$, is estimated at 0.45 and 0.49 under the correct imputation model and under the overfitted imputation model, respectively. However, this percentage is estimated to be much higher under Model 3 (at $\hat{\pi}_{111} \times 100 = 0.82$), and even higher under Model 4 (at $\hat{\pi}_{111} \times 100 = 1.41$) and under Model 5 (at $\hat{\pi}_{111} \times 100 = 1.60$). This result shows that the cell probabilities are estimated more similarly the correct model $\{SC, PC\}$ and under the overfitted model $\{SC, SP, PC\}$. However, there

is a clear difference between $\hat{\pi}_{ijk}$s under Model 3, Model 4 and Model 5 and $\hat{\pi}_{ijk}$ under the correct imputation model $\{SC, PC\}$. The estimated cell probabilities in Table 3.5 can be compared with their ML estimates via the EM algorithm obtained in Little and Rubin (2002, Table 9.9).

Table 3.5. Estimated cell probabilities $\hat{\pi}_{ijk} \times 100$ for candidate imputation models in Survival of infants data. These estimates differ under each model. Although cell probabilities are estimated more similarly under the overfitted model $\{SC, SP, PC\}$ and the correct model $\{SC, PC\}$, there is a clear difference between $\hat{\pi}_{ijk}$s under Model 3, 4 and 5 and $\hat{\pi}_{ijk}$ under the correct imputation model $\{SC, PC\}$.

	Clinic (C)	Prenatal Care (P)	Survival (S) Died	Survived
Model 1: $\{SC, SP, PC\}$	A	Less	0.45	25.35
		More	0.79	38.78
	B	Less	2.64	28.57
		More	0.34	3.07
Model 2: $\{SC, PC\}$	A	Less	0.49	25.27
		More	0.76	38.79
	B	Less	2.68	28.57
		More	0.30	3.15
Model 3: $\{SP, PC\}$	A	Less	0.82	36.66
		More	0.30	28.46
	B	Less	2.27	17.26
		More	0.83	13.40
Model 4: $\{SP, SC\}$	A	Less	1.41	24.64
		More	1.05	38.59
	B	Less	1.68	29.27
		More	0.09	3.23
Model 5: $\{S, P, C\}$	A	Less	1.60	36.34
		More	1.21	27.41
	B	Less	0.81	18.26
		More	0.09	3.27

3.5 Eastern barred bandicoot data revisited

To approach this example, we fitted a logistic regression generalized linear model and assumed that body weights measurements are independent and normally distributed. Also, we let the missingness indicator r_i be a binary variable with the

parameters ϕ_i, $i = 1, \ldots, n$. We applied StEM/improper MI algorithm for handling the missing data under four different imputation model assumptions:

- Model 1: $logit(\phi_i) = \beta_0 + \beta_1 weight_i + \beta_2 gender_i$.
- Model 2: $logit(\phi_i) = \gamma_0 + \gamma_1 weight_i$.
- Model 3: $logit(\phi_i) = \eta_0 + \eta_1 gender_i$.
- Model 4: $logit(\phi_i) = \alpha_0 = const$.

Table 3.6. Information criteria for candidate imputation models in Eastern barred bandicoot data.

	Model 1	Model 2	Model 3	**Model 4**
AIC	2.879	0.986	0.921	**-1.207**
BIC	16.941	12.706	12.640	**8.168**

Table 3.6 shows the results based on the two information criteria AIC and BIC. Here, we set $M = 100$. Based on these results, the MCAR model is the best imputation model among the candidates, that is, missingness occurs due to chance and does neither depend on gender nor on body weight itself, and therefore, may be ignorable. Note that these results might change if we have more information about the data – *i.e.,* if more covariates on individuals were collected in the dataset and were used for modelling.

3.6 Pima Indian Women data revisited

Once again, we fitted a logistic regression generalized linear model assuming three different imputation models. These are

Model 1: the missingness mechanism is nonignorable (MNAR),

Model 2: the missingness mechanism is ignorable (MAR),

Model 3: complete case estimation (MCAR).

Using the StEM/improper MI via Metropolis-within-Gibbs sampler (Tierney, 1999) similar to Section 3.3.2, we assume that under Model 1 the missingness mechanism is $p(r_i \mid X_{mis,i}, X_{obs,i}, Y_i)$, and under Model 2 the missingness mechanism is

$p(r_i \mid X_{obs,i}, Y_i)$ where r_i is the joint missingness indicator and X_i is the joint representation of all explanatory variables. Under the complete case model, we assume that r_i is independent of X_i and Y_i.

Table 3.7. Information criterion for different imputation models fitted to the Pima Indian women data. The data strongly favoured Model 1, suggesting a non-ignorable missing data mechanism.

	Model 1	Model 2	Model 3
AIC	**7231.142**	7660.071	10544.15
BIC	**7714.096**	8026.931	10818.13

Table 3.7 shows the AIC and BIC values compared for each fitted model where both information criteria choose Model 1 over the other models. This strongly suggests that the missingness mechanism is nonignorable. Also, Table 3.8 shows the estimates of regression parameters for each imputation model. Note that results given in this table are based on log-transformations of insulin, pregnancy, pedigree and age as well as standardization on all predictors, and its interpretation here is only used for drawing comparison between the three imputation models. We see that quite different results are obtained for these three models. For example, the regression parameter for 2-hour serum insulin is estimated at 0.321 for Model 1, 0.022 for Model 2 and 0.084 for Model 3. This result suggests how different conclusions can be drawn depending on whether we include the missingness mechanism in the model and whether or not it is worth doing so.

Table 3.8. Estimates of regression parameters and their standard errors (in parentheses) for different imputation models in the Pima Indian women example. Note that pregnancy, insulin, pedigree and age are log-transformed.

	intercept	pregnancy	glucose	pressure	triceps	insulin	bmi	pedigree	age
Model 1	-0.912	0.282	0.918	-0.034	0.163	0.321	0.503	0.328	0.284
	(0.102)	(0.113)	(0.143)	(0.107)	(0.152)	(0.208)	(0.139)	(0.098)	(0.120)
Model 2	-0.890	0.290	1.080	-0.124	0.088	0.022	0.606	0.322	0.315
	(0.100)	(0.113)	(0.185)	(0.110)	(0.199)	(0.267)	(0.179)	(0.098)	(0.119)
Model 3	-1.055	0.076	1.055	-0.025	0.114	0.084	0.449	0.407	0.628
	(0.154)	(0.187)	(0.178)	(0.146)	(0.180)	(0.181)	(0.190)	(0.147)	(0.211)

The nonignorability of the missingness mechanism necessarily relies on parametric assumptions. In this example, we have assumed that $[r_i \mid X_i, Y_i]$ follows a multivariate Bernoulli distribution with a logit link function. Also, for Model 1 we assume nonignorability of missingness mechanism for all the missing variables. A further investigation of different parametric assumptions for r_i (or whether non-

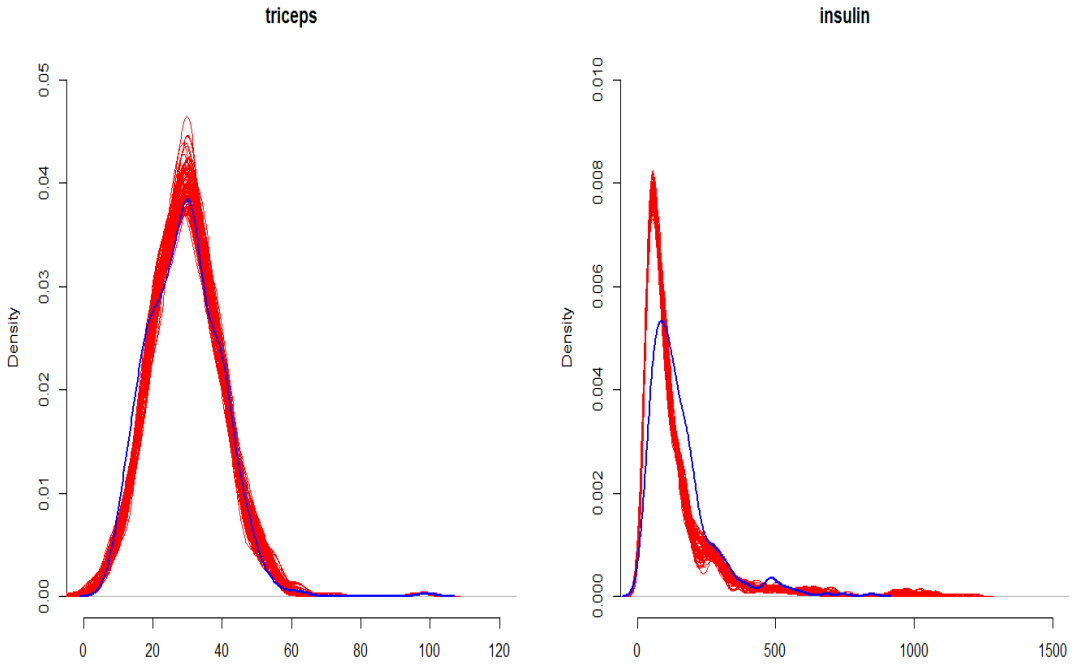


Figure 3.1. Density curve plots of (a) observed triceps skin fold thickness (blue) and multiple imputed values of triceps skin fold thickness (red) and (b) observed 2-hour serum insulin (blue) and multiple imputed values of 2-hour serum insulin (red) for Model 1 in Pima Indian Women example. Similarity between the observed curve and imputed curves in (a) suggests that triceps skin fold thickness might be missing at random. Difference between the observed curve and imputed curves in (b) suggests that 2-hour serum insulin is missing not at random.

ignorability assumption can be relaxed for some of the missing variables) may be carried out if one has a reasonable argument for considering these in the study. For example, a visual inspection of the density plots of the observed and imputed data of triceps skin fold thickness and 2-hour serum insulin for Model 1 in Figure 3.1 might suggest further investigations into whether missingness mechanism for triceps skin fold thickness can be ignored.

3.7 Discussion

In this chapter, we investigated imputation model selection and developed new criteria from methods established in Chapter 2. We investigated whether access to familiar likelihood-based approaches to model selection, such AIC and BIC, allows us to choose the imputation model that best fits the observed data. We tested the performance of the proposed methods on several real data and some simulation studies, and evaluated the post-imputation effect of imputation model

selection on the parameter estimates. We provided insights into imputation model misspecification with the help of variational approximation, and further, examined some theoretical properties of BIC. We showed that BIC can be consistent for the correct imputation model, and that this can even be the case when the correct model is missing not at random.

The fitted models used in this chapter were quite simple so it would be of interest to see how the criteria performs for more complicated scenarios such as in high-dimensional settings where the number of variables or components available for use in the analysis is larger than the number of observations, or when including random effects in the model. We leave these extensions as future work.

3.8 Theoretical arguments: Proof of Theorem 1

Let $y = (y_1, y_2, \ldots, y_{n-m})$ and $z = (z_0, \ldots, z_m)$, where $n \in \mathbb{N}$ and $m \in \{0, 1, \ldots, n-1\}$, and denote $r = (r_1, \ldots, r_n)$ as the missingness indicator. Let $\mathbb{S}$ denote the set of all possible imputation models, $\mathbb{S} = \{\mathbb{M}_k = q(z \mid y, r, \theta_k); k = 0, 1, 2, \ldots, K\}$ with $|\theta_k|$ denoting the total number of parameters for model $\mathbb{M}_k$, and K being the number of competing imputation models. Also, define a subset $\mathbb{M} \subset \mathbb{S}$ of models which minimize $KL(q||p)$. Furthermore, $\mathbb{M}^p \subset \mathbb{M}$ denotes the subset of the most parsimonious correct imputation models, $\mathbb{M}^p = \{\mathbb{M}_k \in \mathbb{M} : |\theta_k| \leq |\theta_{k^*}|, \forall \mathbb{M}_{k^*} \in \mathbb{M}\}$. For simplicity, we denote $q_k(z) = q(z \mid y, r, \theta_k)$ throughout this section.

To prove consistency of BIC for imputation model selection, we require a set of regularity conditions that consist of the following assumptions:

Assumption A1. Observations $z_0, z_1, \ldots, z_m$ are independent, and densities $q_k(z_i)$ exist.

Assumption A2. $m = o_p(n)$ so that the number of observed data points $n - m$ grows faster than the number of missing data m in probability as $n \to \infty$.

Assumption A3. The derivatives of the likelihood function $\int q_k(z) \log\Big(p(y, z, r \mid \theta_k)/q_k(z)\Big)\mathrm{d}z$ up to order three exist *w.r.t.* θ_k, and are continuous and uniformly

bounded for all $\theta_k \in \Theta^{(k)}$.

Assumption A4. The derivatives of observed likelihood function up to order three exist *w.r.t.* θ and are continuous and uniformly bounded for all $\theta \in \Theta$.

We also need Lemmas A.1 and A.2:

Lemma A.1. *Let $q_t(z) \in \mathbb{M}$ be an arbitrary correct imputation model and $q_w(z) \in \mathbb{M}^c$ be an arbitrary wrong imputation model where $\mathbb{M}^c$ denotes the complement of $\mathbb{M}$. Then, we have*

$$\int q_t(z) \log\left(\frac{p(y,z,r \mid \theta_t)}{q_t(z)}\right) \mathrm{d}z > \int q_w(z) \log\left(\frac{p(y,z,r \mid \theta_w)}{q_w(z)}\right) \mathrm{d}z.$$

Proof. The proof is straightforward by using the relationship between the imputation model and observed log-likelihood in (3.1). Suppose $q_k(z)$ is any specified imputation model in $\mathbb{S}$. Let Q be a class of lower bounds based on various imputation models,

$$Q = \left\{ \int q_k(z) \log\left(\frac{p(y,z,r \mid \theta_k)}{q_k(z)}\right) \mathrm{d}z, \ \forall q_k(z) \in \mathbb{S} \right\}.$$

The maximum value of

$$\int q_k(z) \log\left(\frac{p(y,z,r \mid \theta_k)}{q_k(z)}\right) \mathrm{d}z$$

over q_k is obtained when $q_k(z) = p(z \mid y,r,\theta) \in \mathbb{M}$, since the bound from above in (3.1) is attained for $p(z \mid y,r,\theta)$:

$$\int q(z) \log\left(\frac{p(y,z,r \mid \theta)}{q(z)}\right) \mathrm{d}z = \int p(z \mid y,r,\theta) \log\left(\frac{p(y,z,r \mid \theta)}{p(z \mid y,r,\theta)}\right) \mathrm{d}z = \log p(y,r \mid \theta).$$

Therefore, $\forall q_t(z) \in \mathbb{M}$,

$$\int q_t(z) \log\left(\frac{p(y,z,r \mid \theta_t)}{q_t(z)}\right) \mathrm{d}z \ \geq \max_{q_k \in \mathbb{M}^c} \int q_k(z) \log\left(\frac{p(y,z,r \mid \theta_k)}{q_k(z)}\right) \mathrm{d}z.$$

Now, suppose $\exists q_{w^*}(z) \in \mathbb{M}^c$ which maximizes Q over the set of wrong imputation models, and for which the equality holds in the above equation. This would imply

that $q_{w^*}(z) \in \mathbb{M} \notin \mathbb{M}^c$ which would contradict with the assumption of $q_{w^*}(z) \in \mathbb{M}^c$.

Thus, $\forall q_t(z) \in \mathbb{M}$ and $\forall q_w(z) \in \mathbb{M}^c$,

$$\int q_t(z) \log\left(\frac{p(y,z,r \mid \theta_t)}{q_t(z)}\right) \mathrm{d}z \; > \int q_w(z) \log\left(\frac{p(y,z,r \mid \theta_w)}{q_w(z)}\right) \mathrm{d}z.$$

□

Lemma A.2. *Let $q_0(z) \in \mathbb{M}^p$ be the most parsimonious correct imputation model and $q_1(z) \in \mathbb{M}/\mathbb{M}^p$ be an overfitted correct imputation model. Partition*

$$\theta_1 = \begin{bmatrix} \theta_0 \\ \theta_s \end{bmatrix} = \begin{bmatrix} u \times 1 \\ s \times 1 \end{bmatrix}$$

and consider the hypothesis test $H_0 : \theta_s = \mathbf{0}$ vs. $H_1 : \theta_s \neq \mathbf{0}$. Then, under H_0 and as $n \to \infty$,

$$2\left\{\sup_{\theta_1} \int q_1(z) \log\left(\frac{p(y,z,r \mid \theta_1)}{q_1(z)}\right) \mathrm{d}z - \sup_{\theta_0} \int q_0(z) \log\left(\frac{p(y,z,r \mid \theta_0)}{q_0(z)}\right) \mathrm{d}z\right\} \xrightarrow{d} \chi^2_s. \tag{3.3}$$

Proof. Rewrite the null hypothesis as $H_0 : \theta_1 = g(\theta_0) = (\theta_0, \mathbf{0})^\intercal$ such that $G(\theta_0) = \partial g(\theta_0)/\partial\theta_0 = (I_u, \mathbf{0})^\intercal$ where I_u is the $u \times u$ identity matrix and $\mathbf{0}$ is a $s \times u$ matrix of zeros. Let $\hat{\theta}_1$ and $\hat{\theta}_0$ be the MLEs of θ_1 and θ_0, respectively. Also, let $S(\theta)$, $I(\theta)$ and $J(\theta)$ denote the full score function, Fisher and observed information matrix, respectively. Following Sen and Singer (1994, p. 205-207), if Assumptions *A3–A4* are satisfied and as $n \to \infty$, then a Taylor expansion around $\hat{\theta}_1$ yields

$$\begin{aligned} 2\int q_1(z) \log\left(\frac{p(y,z,r \mid \theta_1)}{q_1(z)}\right) \mathrm{d}z = 2\int p(z \mid y,r,\hat{\theta}_1) \log\left(\frac{p(y,z,r \mid \hat{\theta}_1)}{p(z \mid y,r,\hat{\theta}_1)}\right) \mathrm{d}z \\ - S^\intercal(\theta_1)[I(\theta_1)]^{-1} S(\theta_1) + o_p(1), \end{aligned}$$

since $[I(\theta_1)]^{-1} J(\hat{\theta}_1) \xrightarrow{p} I_{u+s}$ (Slutsky's theorem) where I_{u+s} is the identity matrix. Let $S^*(\theta)$, $I^*(\theta)$ and $J^*(\theta)$ denote the restricted score function, Fisher and observed information matrix, respectively. Similarly, since $[I^*(\theta_0)]^{-1} J^*(\hat{\theta}_0) \xrightarrow{p} I_u$,

we may write

$$2\int q_0(z)\log\left(\frac{p(y,z,r\mid\theta_0)}{q_0(z)}\right)\mathrm{d}z = 2\int p(z\mid y,r,\hat{\theta}_0)\log\left(\frac{p(y,z,r\mid\hat{\theta}_0)}{p(z\mid y,r,\hat{\theta}_0)}\right)\mathrm{d}z$$
$$- S^{*\top}(\theta_0)[I^*(\theta_0)]^{-1}S^*(\theta_0) + o_p(1).$$

Now, let $\mathrm{E}_{\theta_k}[.\mid\theta_k]$ be the expectation *w.r.t.* the correct imputation model $p(z\mid y,r,\theta_k)$. Since $\mathrm{E}_{\theta_1}[\log p(y,r\mid\theta_1)] - \mathrm{E}_{\theta_0}[\log p(y,r\mid\theta_0)] = 0$ under H_0, we have

$$2\left\{\int p(z\mid y,r,\hat{\theta}_1)\log\left(\frac{p(y,z,r\mid\hat{\theta}_1)}{p(z\mid y,r,\hat{\theta}_1)}\right)\mathrm{d}z - \int p(z\mid y,r,\hat{\theta}_0)\log\left(\frac{p(y,z,r\mid\hat{\theta}_0)}{p(z\mid y,r,\hat{\theta}_0)}\right)\mathrm{d}z\right\}$$
$$= S^{\top}(\theta_1)[I(\theta_1)]^{-1}S(\theta_1) - S^{*\top}(\theta_0)[I^*(\theta_0)]^{-1}S^*(\theta_0) + o_p(1). \tag{3.4}$$

Note that, under H_0,

$$S^*(\theta_0) = G^{\top}(\theta_0)S(\theta_1).$$

Also, since

$$S^*(\theta_0) \xrightarrow{d} N(0, I^*(\theta_0))$$

and

$$G^{\top}(\theta_0)S(\theta_1) \xrightarrow{d} N(0, G^{\top}(\theta_0)I(\theta_1)G(\theta_0)),$$

under H_0 we may write

$$I^*(\theta_0) = G^{\top}(\theta_0)I(\theta_1)G(\theta_0). \tag{3.5}$$

It follows that (3.4) may be simplified to

$$S^{\top}(\theta_1)\Big[I^{-1}(\theta_1) - G(\theta_0)[I^*(\theta_0)]^{-1}G^{\top}(\theta_0)\Big]S(\theta_1) + o_p(1). \tag{3.6}$$

Furthermore,

$$\begin{aligned} tr\{I^{-1}(\theta_1) - G(\theta_0)[I^*(\theta_0)]^{-1}G^\top(\theta_0)\}I(\theta_1) &= tr\{I_{u+s} - G(\theta_0)[I^*(\theta_0)]^{-1}G^\top(\theta_0)I(\theta_1)\} \\ &= u + s - tr\{[I^*(\theta_0)]^{-1}G(\theta_0)I(\theta_1)G^\top(\theta_0)\} \\ &= u + s - tr\{[I^*(\theta_0)]^{-1}I^*(\theta_0)\} = s \end{aligned} \quad (3.7)$$

Thus, using (3.6) and (3.7), and having $[I^*(\theta_0)]^{-1/2}S^*(\theta_0) \xrightarrow{d} N(0, I_u)$ and $[I(\theta_1)]^{-1/2}S(\theta_1) \xrightarrow{d} N(0, I_{s+u})$ by Slutsky's theorem, we obtain (3.3).

□

Theorem 1. Suppose $\mathbb{M}_0$ is the imputation model chosen by BIC and $\mathbb{M}^p$ is a finite set of the most parsimonious correct models. If Assumptions *A1–A4* are satisfied, then

$$\Pr(\mathbb{M}_0 \in \mathbb{M}^p) \to 1 \;\; as \;\; n \to \infty.$$

Proof. Let $\mathbb{M}_1 \in \mathbb{S}/\mathbb{M}^p$ be an arbitrarily chosen model which is not in the class of the most parsimonious models. To prove Theorem 1, it is sufficient to show that

$$\Pr(\mathrm{BIC}(\mathbb{M}_0) - \mathrm{BIC}(\mathbb{M}_1) < 0) \to 1 \; as \; n \to \infty. \quad (3.8)$$

We have $\mathbb{S} = \mathbb{M} \bigcup \mathbb{M}^c$, where $\mathbb{M}^c$ is the complement of $\mathbb{M}$. Thus, either $\mathbb{M}_1 \in \mathbb{M}/\mathbb{M}^p$ or $\mathbb{M}_1 \in \mathbb{M}^c$. We show that (3.8) holds under both of the following possible cases:

i. $\mathbb{M}_1 \in \mathbb{M}/\mathbb{M}^p$

   We need to show that the logarithmic penalty term in BIC outgrows the difference in the log-likelihood terms as $n \to \infty$. In this case, both $\mathbb{M}_0$ and $\mathbb{M}_1$ are correct, but $\mathbb{M}_1$ is overfitted *w.r.t.* to $\mathbb{M}_0$, that is, $|\theta_1| - |\theta_0| > 0$.

Based on Lemma A.2, we may write

$$
\begin{aligned}
&\lim_{n\to\infty} \Pr\left(\mathrm{BIC}(\mathbb{M}_0) - \mathrm{BIC}(\mathbb{M}_1) < 0\right) \\
&= \lim_{n\to\infty} \Pr\Bigg(2\left\{\sup_{\theta_1} \int q_1(z) \log\left(\frac{p(y,z,r \mid \theta_1)}{q_1(z)}\right) \mathrm{d}z - \sup_{\theta_0} \int q_0(z) \log\left(\frac{p(y,z,r \mid \theta_0)}{q_0(z)}\right) \mathrm{d}z\right\} \\
&\qquad\qquad < \log(n) \times (|\theta_1| - |\theta_0|)\Bigg) \\
&= \lim_{n\to\infty} \Pr\left(\chi^2_{|\theta_1|-|\theta_0|} < \log(n) \times (|\theta_1| - |\theta_0|)\right) = 1.
\end{aligned}
$$

ii. $\mathbb{M}_1 \in \mathbb{M}^c$

Denote $KL(q_k||p)$ as the relative Kullback–Leibler divergence of model $q_k(z)$ from $p(z \mid y, r, \theta_k)$. In this case, $\mathbb{M}_0$ is correct and $\mathbb{M}_1$ is not, that is, $KL(q_1||p) > KL(q_0||p)$. Thus, we need to show that the difference in the log-likelihood terms in BIC outgrows the logarithmic penalty term as $n \to \infty$.

If Assumptions *A1–A2* are satisfied, then

$$
\begin{aligned}
&\frac{1}{2}\left\{\mathrm{BIC}(\mathbb{M}_1) - \mathrm{BIC}(\mathbb{M}_0)\right\} \\
&= -\sup_{\theta_1}\Bigg[(n-m)\left\{\frac{1}{n-m}\int q_1(z)\log\left(\frac{p(y,z,r \mid \theta_1)}{q_1(z)}\right)\mathrm{d}z\right\} \\
&\qquad\qquad + m\left\{\frac{1}{m}\int q_1(z)\log\left(\frac{q_1(z)}{p(z \mid y,r,\theta_1)}\right)\mathrm{d}z\right\}\Bigg] \\
&+ \sup_{\theta_0}\Bigg[(n-m)\left\{\frac{1}{n-m}\int q_0(z)\log\left(\frac{p(y,z,r \mid \theta_0)}{q_0(z)}\right)\mathrm{d}z\right\} \\
&\qquad\qquad + m\left\{\frac{1}{m}\int q_0(z)\log\left(\frac{q_0(z)}{p(z \mid y,r,\theta_0)}\right)\mathrm{d}z\right\}\Bigg] \\
&\qquad\qquad + \frac{1}{2}\log(n) \times (|\theta_1| - |\theta_0|) \\
&= (n-m)\Bigg\{\frac{1}{n-m}\int q_0(z)\log\left(\frac{p(y,z,r \mid \theta_0)}{q_0(z)}\right)\mathrm{d}z \\
&\qquad\qquad - \frac{1}{n-m}\int q_1(z)\log\left(\frac{p(y,z,r \mid \theta_1)}{q_1(z)}\right)\mathrm{d}z\Bigg\}
\end{aligned}
$$

$$
\begin{aligned}
&\quad - m\left\{\frac{1}{m}\int q_1(z)\log\left(\frac{q_1(z)}{p(z\mid y,r,\theta_1)}\right)\mathrm{d}z\right. \\
&\qquad \left. - \frac{1}{m}\int q_0(z)\log\left(\frac{q_0(z)}{p(z\mid y,r,\theta_0)}\right)\mathrm{d}z\right\} + o_p(1) \\
&\qquad + \frac{1}{2}\log(n)\times(|\theta_1| - |\theta_0|) \\
&= (n-m)\left\{\frac{1}{n-m}\int q_0(z)\log\left(\frac{p(y,z,r\mid\theta_0)}{q_0(z)}\right)\mathrm{d}z\right. \\
&\qquad \left. - \frac{1}{n-m}\int q_1(z)\log\left(\frac{p(y,z,r\mid\theta_1)}{q_1(z)}\right)\mathrm{d}z\right\} \\
&\qquad - m\left\{\frac{1}{m}KL(q_1\|p) - \frac{1}{m}KL(q_0\|p)\right\} + o_p(1) \\
&\qquad + \frac{1}{2}\log(n)\times(|\theta_1| - |\theta_0|) \\
&= O_p(n-m) - O_p(m) \pm O(\log(\sqrt{n})) \\
&= O_p(n-m)
\end{aligned}
$$

which tends to be positive since, according to Lemma A.1 and the Law of Large Numbers, the term

$$
(n-m)\left\{\frac{1}{n-m}\int q_0(z)\log\left(\frac{p(y,z,r\mid\theta_0)}{q_0(z)}\right)\mathrm{d}z - \frac{1}{n-m}\int q_1(z)\log\left(\frac{p(y,z,r\mid\theta_1)}{q_1(z)}\right)\mathrm{d}z\right\}
$$

is positive and grows with probability approaching one as $n\to\infty$, with rate $O_p(n-m)$. Thus,

$$
\Pr(\mathrm{BIC}(\mathbb{M}_0) - \mathrm{BIC}(\mathbb{M}_1) < 0) = \Pr(\mathrm{BIC}(\mathbb{M}_1) - \mathrm{BIC}(\mathbb{M}_0) > 0) \to 1 \; as \; n\to\infty.
$$

□

Furthermore, to prove the following corollary, we require the following regularity condition:

Assumption A5. The initial values $\theta^{(0)}$ in the StEM algorithm is consistent asymptotically linear estimate of θ.

Corollary. *If the Assumptions A1–A5 are satisfied, the BIC (see Section 3.2) computed*

at $\theta = \bar{\theta}$ is consistent for imputation model selection.

Proof. Suppose $\hat{\theta}$ is the MLE of θ. It follows that

$$\sqrt{n}(\hat{\theta} - \theta) = n^{-\frac{1}{2}} \sum_{i=1}^{n} [I(\theta)]^{-1} S_i(\theta) + o_p(1) \tag{3.9}$$

as the sample size $n \to \infty$ (Sen and Singer, 1994, p. 205-207).

Also, let $\bar{\theta}$ denote the StEM estimator of θ. Following Wang and Robins (1998, eq. A7) and the regularity condition in Assumption A5, we have

$$\sqrt{n}(\bar{\theta} - \theta) = n^{-\frac{1}{2}} \sum_{i=1}^{n} [I(\theta)]^{-1} S_i(\theta) + o_p(1) \tag{3.10}$$

as $n \to \infty$ and the number of imputations $M \to \infty$.

From (3.9) and (3.10) it follows that

$$\sqrt{n}(\bar{\theta} - \theta) = \sqrt{n}(\hat{\theta} - \theta) + o_p(1) \tag{3.11}$$

as $n \to \infty$ and $M \to \infty$.

Hence, given the Assumptions *A1–A5* and from (3.11), the results of Lemma A.2, and consequently the result of Theorem 1, we may conclude that for sufficiently large number of imputations the BIC obtained from the StEM algorithm is consistent for imputation model selection. □

Chapter 4

Likelihood ratio tests with missing data

4.1 Background

Like model selection, hypothesis testing on model parameters and testing for model goodness-of-fit are essential tools in statistical inference. In the absence of missing data, there are three common test statistics for hypothesis testing of model parameters: likelihood ratio statistic (Neyman and Pearson, 1928), Wald statistic (Wald, 1943), and the score statistic (Rao, 1948). Due to their key role in statistical inference they are sometimes referred to as the “Holy Trinity" in the literature (Rao, 2005). Several studies in statistical inference have reviewed these, either theoretically *e.g.,* Sen and Singer (1994), Casella and Berger (2002) and Boos and Stefanski (2013), illustratively *e.g.,* Buse (1982) and Rayner (1997), or geometrically *e.g.,* Muggeo and Lovison (2014).

The so-called Holy Trinity measures the discrepancy between the null model and sample evidence based on different scales. The likelihood ratio statistic measures the difference between loglikelihoods estimated under the null and alternative models, the Wald statistic works on the parameter scale and measures infeasibility of estimated parameters under the alternative model, and the score statistic works on the score function scale and measures the squared slope at parameters estimated under the restrictions of the null model (Boos and Stefanski, 2013).

Likelihood ratio statistics are also commonly used for checking model goodness-of-fit. Interestingly, these three statistics are asymptotically equivalent (Rao, 1973). However, they differ in second-order properties and may differ in small samples, and thus, possess different advantages and disadvantages. We highlight some of these advantages and disadvantages for cases where data are partially missing, although the main focus of this chapter is on using likelihood ratio statistics.

Wald and score statistics are computationally more efficient than the likelihood ratio statistic. Likelihood ratio statistics are functions of both parameters estimated under the null and alternative models, and in some cases it may be difficult to compute one or the other of these estimates. However, Wald statistics require only parameters under the alternative model to be estimated whereas the score statistic is only a function of estimated parameters under the null model. Thus, the Wald statistic is more convenient when the restricted estimate of θ is difficult to compute and the score statistic is more convenient when the unrestricted estimate of θ is difficult to compute. For example, in situations where significance testing on multiple regression coefficients is of interest – *i.e.,* where they share the same alternative model, then a Wald statistic is computationally preferable because it does not require additional fitting of the null models, and can be computed as a function of only one alternative model (Warton, 2008).

A key drawback of the Wald statistic is its parametrization on the variance (Barndorff-Nielsen and Cox, 1994, p. 120) whereas the likelihood ratio and the score statistic are invariant under reparametrizations. The Wald statistic can perform in an aberrant manner when parameters approach the boundary of a parameter space. For example, Vaeth (1985) showed that the Wald statistic may have poor properties when used with logistic regression models where the power of the test can tend to zero with increasing effect size. Furthermore, reparametrization has a negative effect on the power of Wald statistic when used with generalized estimation equation models, in particular for overdispersed count data (Warton, 2008). A review of the Holy Trinity and a detailed comparison of their advantages and disadvantages are given in Rao (2005).

In the MI literature and in the presence of missing data, methodology for perform-

ing hypothesis tests and checking model goodness-of-fit are not as straightforward to develop. Rubin (1987) proposed to obtain a modified Wald test statistic approximated by the F distribution. However, due to the differences in pros and cons of the Holy Trinity, developing alternative statistics such as the likelihood ratio statistics can be advantageous for small samples. Therefore, subsequent work by Meng and Rubin (1992) developed a pooling procedure for a likelihood ratio test with MI for nested models. They used an approximation based on the asymptotic relationship between the Wald and likelihood ratio statistics. This pooled likelihood ratio statistic was developed as a function of the likelihood ratio statistic averaged across multiple imputed data *and* the likelihood ratio statistic evaluated at $\bar{\theta}$ – *i.e.,* the parameter estimates averaged across multiple imputed data. Although this approach works well for large samples, it can be quite cumbersome to implement in practice due to its multi layered computational requirements. Thus, for example, after reviewing MI's performance in practice, White *et al.* (2011) recommended opting for Rubin's modified Wald statistic for its simplicity.

In this chapter, we focus on likelihood ratio test statistics due to the aforementioned advantages over other test statistics. We develop a likelihood ratio statistic in the presence of missingness and with multiple imputed data and show that asymptotically, it follows a χ^2 distribution for a sufficiently large number of imputations. A theorem is developed to elucidate this result. We achieve this by, once again, exploiting the equivalence between StEM and improper MI. We compare this likelihood ratio statistic, which is obtained from the StEM/improper MI algorithm (see Chapter 2) with the Meng and Rubin (1992) pooled likelihood ratio statistic. We show that a likelihood ratio statistic (when using StEM), is computationally more efficient and outperforms the Meng and Rubin (1992) pooled likelihood ratio statistic for small sample sizes as well as small effect sizes. We demonstrate its numerical performance on simulated and real-data. Following on from previous chapters, our aim is also to show that MI's performance can be improved by accessing likelihood-based tools.

4.2 Likelihood ratio statistic

In this section we provide a brief review of the familiar likelihood ratio statistic in the ML literature, and thereof, develop a likelihood ratio statistic for StEM/improper MI.

4.2.1 Hypothesis testing with MLE

In practice, most of the time we have uncertainty about the true underlying probability distribution of the data, hence, our specified models involve unknown parameters. This uncertainty may be represented by specifying a set of possible models, indexed by a $1 \times s$ parameter vector in the parameter space, $\theta \in \Theta \subset \mathbb{R}^s$, for each hypothesis:

$$H_0 : h(\theta) = 0$$
$$H_1 : h(\theta) \neq 0$$

where $h : \mathbb{R}^s \to \mathbb{R}^u$ is a function such that the $(s \times u)$ matrix of its derivative *w.r.t.* θ exists and is continuous in θ with $\text{rank}[(\partial/\partial\theta)h(\theta)] = u$.

Denote $\Theta_0 = \{\theta : h(\theta) = 0\}$, $\Theta_0 \subseteq \Theta$. An appropriate and well known test statistic for testing H_0 versus H_1 is Wilks' likelihood ratio statistic, given by

$$\Lambda = 2[\sup_{\theta\in\Theta} \ell(\theta) - \sup_{\theta\in\Theta_0} \ell(\theta)].$$

If both supremums are attainable by some estimates, $\hat{\theta}$ and $\hat{\theta}_0$, respectively, then we have

$$\Lambda = 2\{\ell(\hat{\theta}) - \ell(\hat{\theta}_0)\}$$

where the larger Λ is, the more probable H_1 is compared to H_0. If H_0 is true, under certain regularity conditions, Λ follows an asymptotic χ^2_u distribution (Sen and Singer, 1994, Theorem 5.6.1). We consider the above result for cases where data is multiply imputed due to missingness.

4.2.2 Hypothesis testing with multiple imputed data

We draw on the equivalence between StEM and improper MI, discussed in Chapter 2, to develop a likelihood ratio statistic in the presence of missingness and with multiple imputed data. This allows us to conduct (composite) hypothesis tests and check for the goodness-of-fit for models of interest. The key result presented in this chapter is given in the following theorem for which we give a proof in Section 4.6:

Theorem 2. *Let $\theta \in \Theta \subset \mathbb{R}^s$. Consider the hypothesis test*

$$\begin{aligned} H_0 &: h(\theta) = 0 \\ H_1 &: h(\theta) \neq 0 \end{aligned} \tag{4.1}$$

where $h : \mathbb{R}^s \to \mathbb{R}^u$ is a function such that the $(s \times u)$ matrix of its derivative w.r.t. θ exists and is continuous in θ with $\operatorname{rank}[(\partial/\partial\theta)h(\theta)] = u$. Let $\omega \in \mathbb{R}^{s-u} = \{\theta \in \Theta : h(\theta) = 0\}$. Denote $\bar{\theta}$ and $\bar{\omega}$ as the StEM estimators of θ and ω, respectively. If Assumptions B1–B4 (see Section 4.6) are satisfied, the likelihood ratio statistic for the hypothesis test of interest is

$$\Lambda_{st}(y, r) = 2\left\{\ell(\bar{\theta}; y, r) - \ell(\bar{\omega}; y, r)\right\},$$

and its asymptotic distribution is χ^2_u under H_0.

4.3 Simulation study

To assess the performance of the likelihood ratio statistic under StEM and to numerically investigate Theorem 2, we conducted a simulation study. The focus of this simulation study is on hypothesis testing and examining statistical power under various parameter settings as well as for various sample sizes.

In this simulation study we investigated the performance of the pooled likelihood ratio statistic by MI (throughout, we denote this as the Pooled LR, Meng and Rubin, 1992) and the likelihood ratio statistic obtained from the StEM algorithm (denoted by StEM LRT, see Section 4.2). For simplicity, we constructed a linear regression model simulation where only one predictor was used, however, the

following simulation study can easily be extended to multiple predictors.

Let $X_i \sim N(0,1)$, $i = 1,\ldots,n$, be the predictor which is fully observed, and Y_i be the response variable, where $\mathrm{E}(Y_i \mid X_i, \beta) = \beta_0 + \beta_1 X_i$. Suppose that Y_i is partially observed with probability

$$p(r_i = 1 \mid Y_i, X_i, \gamma) = \frac{\exp(\gamma_0 + \gamma_1 X_i)}{1 + \exp(\gamma_0 + \gamma_1 X_i)}$$

of Y_i being observed, where r_i denotes the missingness indicator. We set the sample size n to be between 30 and 1000, with $\beta_0 = -1$, and various values for the slope $\beta_1 \in [-1, 1]$. Also, we set $\gamma = (-1, -0.1)^\intercal$ to achieve approximately 25% missing proportion in Y_i. The number of multiple imputations was set to $M = 100$, see Section 2.4.2 for further details on selecting M for StEM. We then considered the following two models:

Model 1: $\mathrm{E}(Y_i \mid X_i, \beta) = \beta_0 + \beta_1 X_i$,

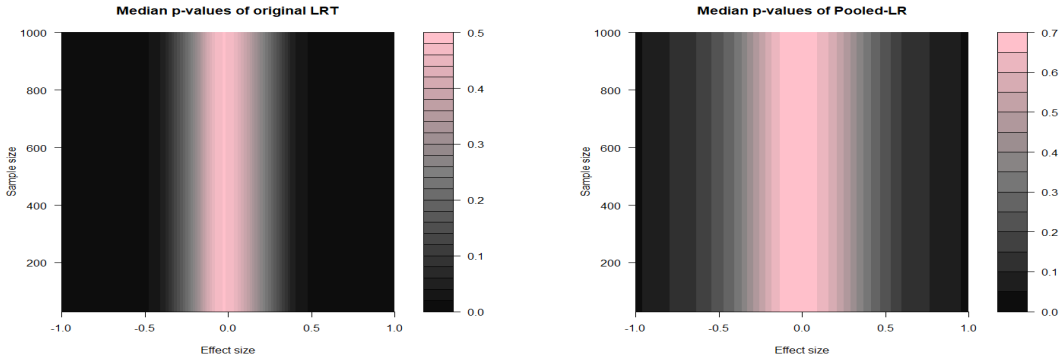
Model 0: $\mathrm{E}(Y_i \mid X_i, \beta) = \beta_0$,

such that Model 0 is nested within Model 1. Under this setting, we considered the following hypothesis test:

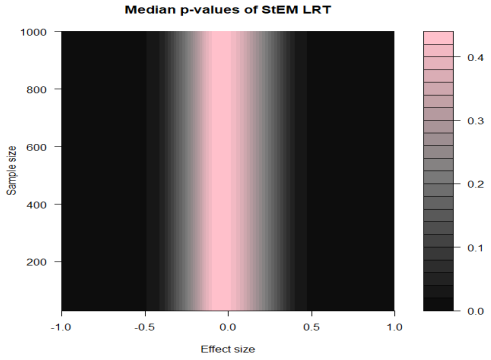
$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0.$$

We ran 500 simulations for each combination of n and β_1, and compared the performance of the LRT using the original (complete) dataset with the StEM LRT and the Pooled LR based on p-values and statistical power of tests. Figure 4.1 shows the median p-values of each method for different values of β_1 and n. The true effect size (β_1) is on the x-axis and ranges from -1 to +1. The sample size (n) is on the y-axis and ranges from 30 to 1000. Furthermore, the shades of colour from dark grey to pink show the median p-values of 0 and above: the closer the median p-value to 0 the darker the shade. For $\beta_1 = 0$, we would expect the average p-values to be 0.5, thus we would expect to see pink shades when $\beta_1 = 0$ and as β_1 shifts away from 0 we would expect to see darker greys. Ideally, this will show a close resemblance to the shades (median p-values) of the original

LRT in Figure 4.1a.We see that the StEM LRT in Figure 4.1c performed similarly to the LRT based on the original dataset, whereas the Pooled LR in Figure 4.1b performed poorly for small effect sizes even when the sample size was moderate to large (pink/light grey shade). Also, there is more spread in the Pooled LR's p-values ranging from 0 to 0.7 compared to StEM LRT's 0 to 0.48. The latter are closer to the original LRT's 0 to 0.5, and both increase more sharply as the effect size decreases.



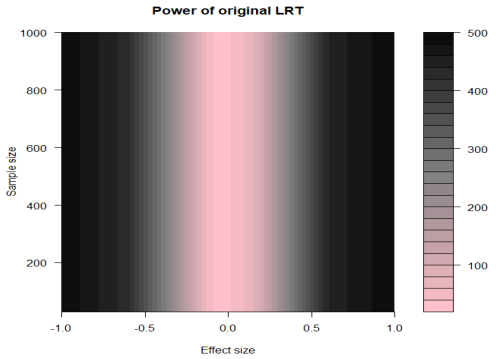
(a) (b)



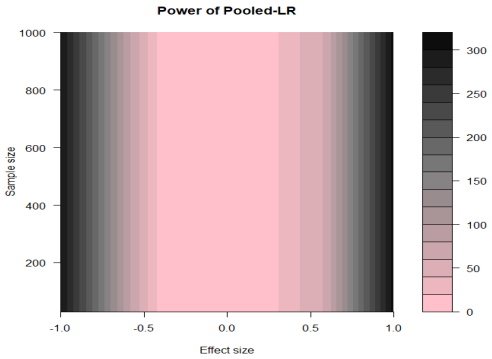
(c)

Figure 4.1. Median p-values of the tests in 500 simulated datasets. (a) shows median p-values of LRT based on original (complete) datasets. (b) shows median p-values of Pooled LR based on the MI algorithm. (c) shows median p-values of LRT based on the StEM algorithm.

Furthermore, based on the power of the test at significance level of 0.05, StEM outperformed MI as is shown in Figure 4.2. The power of the StEM LRT ranged from 0.12 to 0.992 and again the test performed more similarly to the original LRT's 0.07 to 0.998. The power of Pooled LR ranged from 0 to 0.614 and the test performed poorly particularly for small sample sizes even when the effect size is large; as well as for small effect sizes even when the sample size is large (pink shade). Our simulation results show that StEM LRT outperforms Pooled LR for small samples as well as for small effect sizes in terms of both the average p-values and the statistical power of the test.

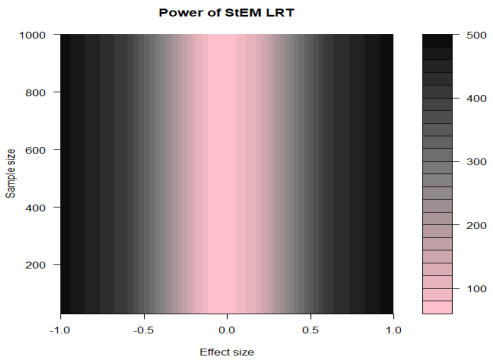


(a)



(b)

Figure 4.2. Power of the tests in 500 simulated datasets at a significance level of 0.05. (a) shows power of LRT based on original (complete) datasets. (b) shows power of Pooled LR based on the MI algorithm. (c) shows power of LRT based on the StEM algorithm.



(c)

4.4 Survival of infants data revisited

Next, we explored the usefulness of the likelihood ratio statistic under StEM using the survival of infants data, see Section 3.2.1 for further details. We compared the null models $\{SC, SP, PC\}$, $\{SC, PC\}$, $\{SP, SC\}$ and $\{S, P, C\}$ against the saturated model $\{SCP\}$ using the StEM LRT discussed in Section 4.2. Recall that this example was analyzed in Little and Rubin (2002, p. 192) and Meng and Rubin (1992) based on likelihood ratio obtained from the EM algorithm and Pooled LR, respectively, where both studies shared the same conclusion that $\{SC, PC\}$ is the preferred parsimonious fitted model.

Table 4.1 shows the value of likelihood ratio statistic and the p-value of the StEM LRT for each null model. Based on these results, we fail to reject the null models $\{SC, SP, PC\}$ ($p-\text{value} = 0.396 > 0.1$) and $\{SC, PC\}$ ($p-\text{value} = 0.556 > 0.1$) against the saturated model $\{SCP\}$. Hence, we can conclude that the conditional independence model given Clinic *i.e.,* $\{SC, PC\}$, is the preferred parsimonious fitted model. This conclusion based on our proposed likelihood ratio test method is

consistent with those of Little and Rubin (2002) and Meng and Rubin (1992).

Table 4.1. Likelihood ratio test for different null models against the saturated Model $\{SCP\}$ in the survival of infants example.

	$\{SC, SP, PC\}$	$\{\mathbf{SC}, \mathbf{PC}\}$	$\{SP, SC\}$	$\{S, P, C\}$
Δ_{st}	0.72	**1.17**	132.32	151.24
p-value	0.396	**0.556**	0.000	0.000

4.5 Discussion

As discussed earlier, hypothesis testing for model parameters when using multiple imputed data is more difficult compared with the complete data case. However, since the likelihood function is available to us through the StEM framework, we found that producing theoretical results (that is, Theorem 2) for the likelihood ratio statistic was a lot easier compared with Wald or score test statistics.

The StEM likelihood ratio statistic is easy to implement and our simulation study shows that it performs well for small samples as well as for small effect sizes. This approach may also be applied in a wide range of situations where Wilk's likelihood ratio statistic is applicable *e.g.,* to evaluate goodness-of-fit of a model of counts when comparing Poisson and negative binomial models, or, to test for the order of a finite Markov chain (Anderson and Goodman, 1957).

In the same spirit as in this chapter, investigating other extensions such as developing a score test statistic using StEM may be possible. As discussed earlier, score statistics have some nice advantages such as computational efficiency that makes them attractive when conducting hypothesis tests and testing for goodness-of-fit, for example, in the context of generalized linear models to test for adequacy of the link function (Pregibon, 1980), overdispersion (Dean, 1992), or zero-inflation (Deng and Paul, 2000). Score statistics can also handle multivariate correlated data well, see Shen and Chen (2012) for cases involving drop-out missingness and Stoklosa *et al.* (2014) for cases where there are many predictor variables.

Furthermore, there is little work in the MI literature about hypothesis testing and

checking model goodness-of-fit in high dimensions where the number of variables may largely exceed the number of observations. However, this is an active area of research in the likelihood-based literature. For example, recently Shah and Bühlmann (2018) have developed a model goodness-of-fit checking tool for linear high dimensional data that is based on residual prediction test and is amenable to a wide range of prediction methods such as lasso, ridge regression, elastic net and ordinary least squares (Hastie *et al.*, 2015). Again, it is possible to implement and access likelihood-based tools such as these, which can improve MI's performance in goodness-of-fit tests for high dimensional data. We leave these extensions as future work.

4.6 Theoretical arguments: Proof of Theorem 2

Let $\ell(\theta; y, r) = \log p(y, r \mid \theta)$ be the observed log-likelihood function, $S(\theta)$ be the score function of the observed log-likelihood, $I(\theta)$ be the Fisher information matrix, and $J(\theta)$ be the observed information matrix. MLEs are solutions to $S(\theta) = 0$.

To prove Theorem 2, we require the following regularity conditions:

Assumption B1. The derivatives of observed likelihood up to order three exist *w.r.t.* θ and are continuous and bounded for all $\theta \in \Theta$.

Assumption B2. The Fisher information matrix, $I(\theta)$, is finite and positive definite.

Assumption B3. For every ω in the closure of a neighbourhood of θ, $I^*(\omega)$ is finite and positive definite.

Assumption B4. The initial values $\theta^{(0)}$ and $\omega^{(0)}$ in the StEM algorithm are consistent asymptotically linear estimates of θ and ω, respectively.

Proof. First, following along the lines of Sen and Singer (1994, p. 239), hypothesis test (4.1) may be written as

$$H_0 : \theta = g(\omega)$$
$$H_1 : \theta \neq g(\omega)$$

where $\omega \in \mathbb{R}^{s-u}$, and $g : \mathbb{R}^{s-u} \to \mathbb{R}^s$ such that the $(s \times s-u)$ matrix of its derivative *w.r.t.* ω exists with

$$\text{rank}[(\partial/\partial\omega)g(\omega)] = s - u.$$

Suppose $\hat{\theta}$ is the full MLE of θ and $\hat{\omega}$ the MLE of ω. It follows that

$$\sqrt{n}(\hat{\theta} - \theta) = n^{-\frac{1}{2}} \sum_{i=1}^{n} [I(\theta)]^{-1} S_i(\theta) + o_p(1) \tag{4.2}$$

as the sample size $n \to \infty$ (Sen and Singer, 1994). Similarly, as $n \to \infty$ we have

$$\sqrt{n}(\hat{\omega} - \omega) = n^{-\frac{1}{2}} \sum_{i=1}^{n} [I^*(\omega)]^{-1} S_i^*(\omega) + o_p(1) \tag{4.3}$$

where $S^*(\omega) = n^{-\frac{1}{2}} \partial \ell(\omega; y, r)/\partial\omega$ and $I^*(\omega) = \mathrm{E}\{\partial^2 \ell(\omega; y, r)/\partial\omega\partial\omega^\top\}$.

Also, denote $\bar{\theta}$ as the full StEM estimator of θ and $\bar{\omega}$ the StEM estimator of ω. Following Wang and Robins (1998) and the regularity condition in Assumption B4, we have

$$\sqrt{n}(\bar{\theta} - \theta) = n^{-\frac{1}{2}} \sum_{i=1}^{n} [I(\theta)]^{-1} S_i(\theta) + o_p(1) \tag{4.4}$$

as $n \to \infty$ and the number of imputations $M \to \infty$. Similarly, we may write

$$\sqrt{n}(\bar{\omega} - \omega) = n^{-\frac{1}{2}} \sum_{i=1}^{n} [I^*(\omega)]^{-1} S_i^*(\omega) + o_p(1) \tag{4.5}$$

as $n \to \infty$ and $M \to \infty$.

From (4.2) and (4.4) it follows that

$$\sqrt{n}(\bar{\theta} - \theta) = \sqrt{n}(\hat{\theta} - \theta) + o_p(1) \tag{4.6}$$

as $n \to \infty$ and $M \to \infty$. Similarly, from (4.3) and (4.5) it follows that

$$\sqrt{n}(\bar{\omega} - \omega) = \sqrt{n}(\hat{\omega} - \omega) + o_p(1) \tag{4.7}$$

as $n \to \infty$ and $M \to \infty$.

Now, consider the following second-order Taylor expansion

$$
\begin{aligned}
2\ell(\bar{\theta};y,r) &= 2\ell(\hat{\theta};y,r) + (\hat{\theta}-\bar{\theta})^\top \frac{\partial^2}{\partial\theta\partial\theta^\top}\ell(\theta;y,r)|_{\theta=\hat{\theta}}\ (\hat{\theta}-\bar{\theta}) + o_p(1)\\
&= 2\ell(\hat{\theta};y,r) - n(\hat{\theta}-\bar{\theta})^\top J(\hat{\theta})(\hat{\theta}-\bar{\theta}) + o_p(1)\\
&= 2\ell(\hat{\theta};y,r) - \sqrt{n}\left\{(\hat{\theta}-\theta)-(\bar{\theta}-\theta)\right\}^\top I(\theta)[I(\theta)]^{-1}J(\hat{\theta})\ \sqrt{n}\left\{(\hat{\theta}-\theta)-(\bar{\theta}-\theta)\right\}\\
&\quad + o_p(1)\\
&= 2\ell(\hat{\theta};y,r) + o_p(1)\ I(\theta)\ o_p(1) + o_p(1)\\
&= 2\ell(\hat{\theta};y,r) + o_p(1)
\end{aligned}
$$

by (4.6) and since $[I(\theta)]^{-1}J(\hat{\theta}) \xrightarrow{p} I_s$ (Slutsky's theorem) where I_s is the $(s\times s)$ identity matrix.

Similarly, using a second-order Taylor expansion around $\hat{\omega}$ we have

$$
\begin{aligned}
2\ell(\bar{\omega};y,r) &= 2\ell(\hat{\omega};y,r) + (\hat{\omega}-\bar{\omega})^\top \frac{\partial^2}{\partial\omega\partial\omega^\top}\ell(\omega;y,r)|_{\omega=\hat{\omega}}\ (\hat{\omega}-\bar{\omega}) + o_p(1)\\
&= 2\ell(\hat{\omega};y,r) - n(\hat{\omega}-\bar{\omega})^\top J^*(\hat{\omega})(\hat{\omega}-\bar{\omega}) + o_p(1)\\
&= 2\ell(\hat{\omega};y,r) - \sqrt{n}\left\{(\hat{\omega}-\omega)-(\bar{\omega}-\omega)\right\}^\top I^*(\omega)[I^*(\omega)]^{-1}J^*(\hat{\omega})\ \sqrt{n}\left\{(\hat{\omega}-\omega)-(\bar{\omega}-\omega)\right\}\\
&\quad + o_p(1)\\
&= 2\ell(\hat{\omega};y,r) + o_p(1)
\end{aligned}
$$

by (4.7) and since $[I^*(\omega)]^{-1}J^*(\hat{\omega}) \xrightarrow{p} I_{s-u}$ (Slutsky's theorem) with I_{s-u} being the $(s-u\times s-u)$ identity matrix, and where $J^*(\omega) = -n^{-1}\partial^2\ell(\omega;y,r)/\partial\omega\partial\omega^\top$ is the observed information matrix under H_0.

Hence,

$$
\begin{aligned}
\Lambda_{st}(y,r) &= 2\left\{\ell(\bar{\theta};y,r) - \ell(\bar{\omega};y,r)\right\}\\
&= 2\left\{\ell(\hat{\theta};y,r) - \ell(\hat{\omega};y,r)\right\} + o_p(1)\\
&= \Lambda(y,r) + o_p(1),
\end{aligned}
$$

as n and M both approach ∞, with $\Lambda(y,r) = 2\left\{\ell(\hat{\theta};y,r) - \ell(\hat{\omega};y,r)\right\}$ being the log-likelihood ratio statistic computed at the MLEs.

Finally, since $\Lambda(y,r) \xrightarrow{d} \chi^2_u$ (Wilks, 1938; Chernoff, 1954; Sen and Singer, 1994), we may write

$$\Lambda_{st}(y,r) \xrightarrow{d} \chi^2_u$$

under H_0 as $n \to \infty$ and $M \to \infty$.

□

Chapter 5

Measurement error modelling with missing data in covariates

5.1 Background

This chapter is designed to address the problem of parameter estimation when explanatory variables are subject to measurement error and missingness. We primarily focus on computational aspects rather than theory. Also, unlike previous chapters where missingness can occur in the response or in the explanatory variable, here, the missingness can only be in the explanatory variable. Borrowing ideas from Chapter 2, the aim of this chapter is three-fold: (1) we propose a new approach to deal with measurement error and missingness, (2) examine model performance through simulation studies and a real example, and (3) demonstrate the simplicity in fitting these methods (*i.e.,* by using existing and well-known software) and examining computational efficiency by comparing computational times.

The combined presence of missingness and measurement error in explanatory variables may have a double effect on statistical analyses when not accounted for. On the one hand, the presence of missing data, in particular, when the missingness does not arise due to chance, can complicate statistical analyses as many methods are primarily designed for complete datasets, see Section 1.3 for further details on the impacts of missing data on inference. Moreover, numerous studies have shown that ignoring missing observations, particularly in explanatory variables, may re-

sult in bias in estimates of regression coefficients and loss of power (Schafer, 1999; Little and Rubin, 2002; Arunajadai and Rauh, 2012). On the other hand, uncertainty in explanatory variables due to measurement error may also result in bias in estimates of regression coefficients and loss of power; a phenomena that is referred to in the literature as the *double whammy* (Carroll *et al.*, 2006). To avoid misleading conclusions and poor inference, it is therefore of high importance to correct for both.

Methods developed specifically for handling missing data or measurement error in explanatory variables have been extensively addressed in the literature, however, fewer studies have addressed these simultaneously. Motivated by an example in nutritional epidemiology, Carroll *et al.* (1997) compared model robustness of maximum likelihood estimation with method of moments when explanatory variables are both subject to missingness and measurement error. Wang *et al.* (2008) proposed an approach based on expected estimation equations as a unified solution to missingness, errors-in-variables and missclassification in explanatory variables, and studied their asymptotic properties. In order to fit these models however one requires either calibration data or a validation subset, which may not always be available. Yi *et al.* (2012) considered longitudinal data and developed a corrected score function with inverse probability weights (IPW) to incorporate measurement error and missingness effects, respectively, followed by the generalized method of moments to combine their results. Although IPW is commonly applied to missing data problems, the method is sensitive to outliers which can result in very large weights and obscure inference, see Seaman and White (2011). Finally, Shen and Chen (2016) used generalized method of moments to handle both measurement error and missing values for generalized linear models (GLMs) but the focus there was on model selection of explanatory variables.

As discussed in Section 1.5.3 there are two types of measurement error models that incorporate uncertainty in explanatory variables in the analyses: functional or structural. For functional measurement error models, the distribution of the true covariate need not be specified whereas structural measurement error models rely on distributional assumptions for the true covariate. Therefore, functional

measurement models are advantageous over structural types in many practical situations where there is little knowledge available about the distribution of the true covariate. Two well-known functional measurement error models that account for measurement errors in explanatory variables in the analyses are *Simulation extrapolation* and *Corrected score*. In this thesis we specifically focus on these two methods but we note that there exists many measurement error methods that could similarly be used, see Carroll *et al.* (2006) for more details. We chose these methods due to their simplicity, convenient form (in the sense that these methods are directly applicable with the approach presented in Chapter 2) and their accompanying software.

Simulation extrapolation (SIMEX, Cook and Stefanski, 1994; Stefanski and Cook, 1995) is a Monte Carlo approach used for finding a relationship between the measurement error variance and the measurement error-induced bias in order to estimate and correct for this bias. Although SIMEX was originally designed for continuous explanatory variables, it was later developed for discrete explanatory variables (Gustafson, 2003; Küchenhof *et al.*, 2006). SIMEX is a general methodology in the sense that it can be applied to almost any measurement error model and results in approximately unbiased and consistent estimates of regression parameters (Carroll *et al.*, 2006). A key advantage in using SIMEX is its simplicity in fitting models when the response is non-normal (*i.e.,* for a known value of the measurement error variance, GLMs with a parametric form can be flexibly and easily fitted).

Corrected score (CS, Stefanski, 1989; Nakamura, 1990) is an alternative functional approach that is based on finding the so-called "corrected score" as a differential of a corrected log-likelihood function. For example, Nakamura (1990) and Stefanski (1989) identified corrected scores for certain models such as Poisson and gamma regression models. Further extensions to CS include the weighted corrected score to fit logistic regression models (Chen *et al.*, 2015), rare-event logistic and extreme-value binary regression models (Buzas and Stefanski, 1996), and stochastic version of CS for approximating corrected scores based on Monte Carlo averaging (Novick and Stefanski, 2002). For a detailed discussion on CS

(which details calculation of standard errors, etc.) see Carroll *et al.* (2006). Generally, CS is simple to program and is computationally faster than SIMEX.

In this chapter, we use the connection between MI and StEM (see Section 2.4) and combine the improper MI/StEM approach with SIMEX and CS to deal with the combined problem of missingness and measurement error in explanatory variables. We give a brief overview of the two measurement error models mentioned above throughout Section 5.2.1. A heuristic explanation of the proposed combined methods is given in section 5.2.2. We then investigate their performance in two simulation studies in Section 5.3, and apply the methods to real data in Section 5.4. Finally, we provide a summary discussion in Section 5.5.

5.2 Methodology

5.2.1 Available functional methods

In this section, we discuss two statistical approaches that reflect uncertainty in the explanatory variables due to measurement errors. As mentioned in Section 1.5.3, the underlying error in the explanatory variables can be classified into classical error and Berkson error (Fuller, 1987; Carroll *et al.*, 2006). Throughout this chapter, we focus on classical error, however, the methods presented in the next section can be applied to Berkson error as well.

A classical measurement error model assumes that the observed explanatory variable (W) captures the true explanatory variable (X) with some additive noise, that is, $W = X + U$, where U is the measurement error with mean 0 and variance $\sigma_U^2 > 0$ and is independent from X and the response variable Y. Naïve estimators of model parameters which substitute W with X without making any adjustments for this substitution are often inconsistent (Armstrong, 1985; Stefanski and Carroll, 1985; Fuller, 1987) and would result in bias (Stefanski and Carroll, 1985; Penev and Raykov, 1993; Carroll *et al.*, 2006) for large σ_U^2.

Simulation extrapolation (SIMEX). The basic idea of SIMEX is to experimentally find the effect of measurement error on an estimator via a reasonably large number of simulations in order to correct for this effect without the need for model-fitting the error (Carroll *et al.*, 2006). The SIMEX estimation proceeds as follows.

Suppose that $U \sim N(0, \sigma_U^2)$ where σ_U^2 is known or can be reasonably well estimated from auxiliary data or a validation dataset. Consider an arbitrary sequence $0 = \lambda_1 < \cdots < \lambda_k < \cdots < \lambda_K$. For each λ_k, create a new dataset by adding inflated errors $U^* \sim N(0, \lambda_k \sigma_U^2)$ to the observed explanatory variable (W). This creates additional datasets with increasingly larger measurement error variance $\sigma_U^2 + \lambda_k \sigma_U^2$. Using these new datasets, estimate the model parameters. Repeat the two steps of simulation and estimation a large number of times, say B times. Obtain the Monte Carlo average of estimates of model parameters over B simulated datasets. Plot these averaged, error contaminated estimates against λ_k, $k = 1, \ldots, K$, in order to find a model that represents the relationship between the two. Finally, since the true error variance is equal to $\sigma_U^2 + \lambda_k \sigma_U^2$, the SIMEX estimation is achieved by extrapolating back to $\lambda = -1$ which is the ideal case of no errors.

What makes SIMEX attractive is the fact that it is very simple to implement, in particular, due to the current software readily available, such as `R` (using the `simex` `R`-package, see Lederer and Kuchenhoff, 2006) and `STATA`, to fit generalized linear models. However, it is computationally intensive and could be computationally inefficient under certain circumstances such as in high dimensional data settings. Also, care must be given to the extrapolant function in complicated situations such as when measurement errors are correlated. For a detailed discussion on SIMEX see Carroll *et al.* (2006) and the references therein.

Corrected score (CS). An alternative approach for dealing with errors-in-variables is to correct for the effects of contaminated explanatory variables via the score function. The key idea underlying CS is to take advantage of the fact that the conditional distribution of an unbiased estimator, such as MLE, given the true observed variables is centered around the true parameter values and thus, to center

the conditional distribution of the corrected estimator given the measurement error around the unbiased estimator (Stefanski, 1989; Nakamura, 1990).

The CS estimation proceeds as follows. Recall that $U \sim N(0, \sigma_U^2)$. Let θ and $S(\theta; X, Y)$ denote the model parameters and an unbiased measurement error-free score function, respectively. Since X is not observed, $S(\theta; X, Y)$ cannot be used for estimation. Thus, with a corrected score function we opt to obtain an unbiased estimator of $S(\theta; X, Y)$ based on the observed data. A corrected score function, $S^*(\theta; W, Y)$, is a function whose conditional expectation *w.r.t.* the measurement errors, $\mathrm{E}[S^*(\theta; W, Y) \mid Y, X]$, coincides with the true score function, $S(\theta; X, Y)$, for all Y, X and θ. The CS estimate is a solution to $S^*(\theta; W, Y) = 0$, see Nakamura (1990).

CS can be used for parameter estimation as well as for inference. Also, under mild regularity conditions, CS results in fully consistent estimates as opposed to SIMEX whose estimator is only approximately consistent in many complex cases such as in logistic regression models (Chen *et al.*, 2015). Furthermore, CS is simple to program and is computationally fast. A drawback for using CS is that it is not always easy to identify a corrected score.

5.2.2 Functional methods with multiple imputation

To simultaneously account for both missingness and measurement error in explanatory variables, we combine the improper MI/StEM approach, with the above measurement error models. In this method, we first apply MI to create multiple imputed datasets. Next, we implement a measurement error model either SIMEX (MI-SIMEX) or CS (MI-CS) to each imputed dataset. Finally, the results, *i.e.,* SIMEX estimates or CS estimates, are combined over multiple imputed datasets following MI's combination rules. Since we propose to apply an improper version of MI (as in Chapter 2), MI's combination rules need to be modified for *improper* MI as shown in Diebolt and Ip (1996). Here, the point estimates of MI-CS and MI-SIMEX are the CS and SIMEX estimates averaged over multiple imputed datasets, respectively. Also, based on the Louis' method (Louis, 1982), their auxiliary esti-

mates are derived from the difference between the information matrices of the CS and SIMEX estimates averaged over multiple imputed datasets *and* the covariance of their respected score functions between multiple imputed datasets.

The choice of a measurement error modelling method can proceed based on their pros and cons available in the measurement error literature, that is, as if there were no missing data. MI-SIMEX is easy to implement since software is currently available to perform SIMEX for generalized linear models. However, MI-CS is computationally more efficient since it does not require large amounts of resampling as in SIMEX. No calibration data or a validation subset is required for fitting each of these methods.

5.3 Simulation study

To investigate the performances of MI-SIMEX and MI-CS in handling both the missingness and errors-in-variables in explanatory variables, we conducted two simulation studies where (1) the response variable is normal; and (2) the response variable is non-normal. These two simulation studies and their results are discussed in the following sections.

5.3.1 Normal linear model

First, we consider the following linear regression model,

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \quad \epsilon_i \overset{i.i.d.}{\sim} N(0,1), \quad i = 1, \ldots, n,$$

where the partially observed and error-contaminated explanatory variable (X_i) was generated under two scenarios: from (a) the normal distribution with mean 0 and variance 1 and (b) the uniform distribution on the interval $(-1, 1)$. The latter investigates the case where the distribution of the true covariate is non-normal. We assumed the error (U_i) in the explanatory variable to be normally distributed with mean 0 and variance σ_U^2. In addition, we assumed X_i to be missing at random where the probability of missingness was generated from a logistic model condi-

tional on the response variable only. To see an example on how we simulated missing data, see Section 3.3 in the error free case.

By considering the model structure above, and using both scenarios for X_i, $i = 1, \ldots, n$, we evaluated model performance based on the magnitude of bias and the mean squared error (MSE) of the slope coefficient (β_1) for different values of $\sigma_U \in \{0.1, 0.25, 0.5\}$ as well as for various missing proportions for the explanatory variable values, set to 25%, 35% and 45%. We set the true regression coefficients to $\beta = (\beta_0, \beta_1)^T = (-0.5, 1)^T$. For each scenario, we then generated the response variable Y_i, $i = 1, \ldots, n$, added measurement error to X_i and imposed missingness on it. We fitted the naïve model and the four linear regression MI models (see Chapter 2), SIMEX (see Section 5.2.1), MI-SIMEX and MI-CS (see Section 5.2.2) The naïve model refers to the complete-case analysis where neither missingness nor measurement error are accounted for.

We performed 100 simulations under each scenario for various sample sizes over a range of values of $n \in \{50, 100, 1000\}$. In addition, throughout this section, we use the results of the fully observed and error-free dataset (complete) as a baseline measurement in the visual inspection of the above mentioned compared methods.

We used the `mice` R-package (Buuren and Groothuis-Oudshoorn, 2011) for multiple imputation using the packages default settings; the `simex` R-package when fitting SIMEX models; and wrote our own code for fitting CS models. Finally, for MI we set $M = 100$ (see Section 2.3.2), and for SIMEX we set $B = 100$, $\lambda_k = \{0.25, 0.5, 1, 1.5\}$ and used a quadratic function for the extrapolation step.

In Figure 5.1 (and Figures A.1–A.2 in Appendix) and Figure 5.2 (and Figures A.3–A.4 in Appendix), we plotted parallel boxplots of the estimates of the slope coefficient obtained from each fitted model against increasing missing proportions and increasing values of σ_U for various n under the first and the second scenarios, respectively. For example, for $n = 100$, Figure 5.1 and Figure 5.2 show that MI, MI-SIMEX and MI-CS outperformed the other methods in terms of the magnitude of bias as the missing proportion increases (top to bottom). However, as σ_U increases (from left to right) MI fails to perform well compared to MI-SIMEX and

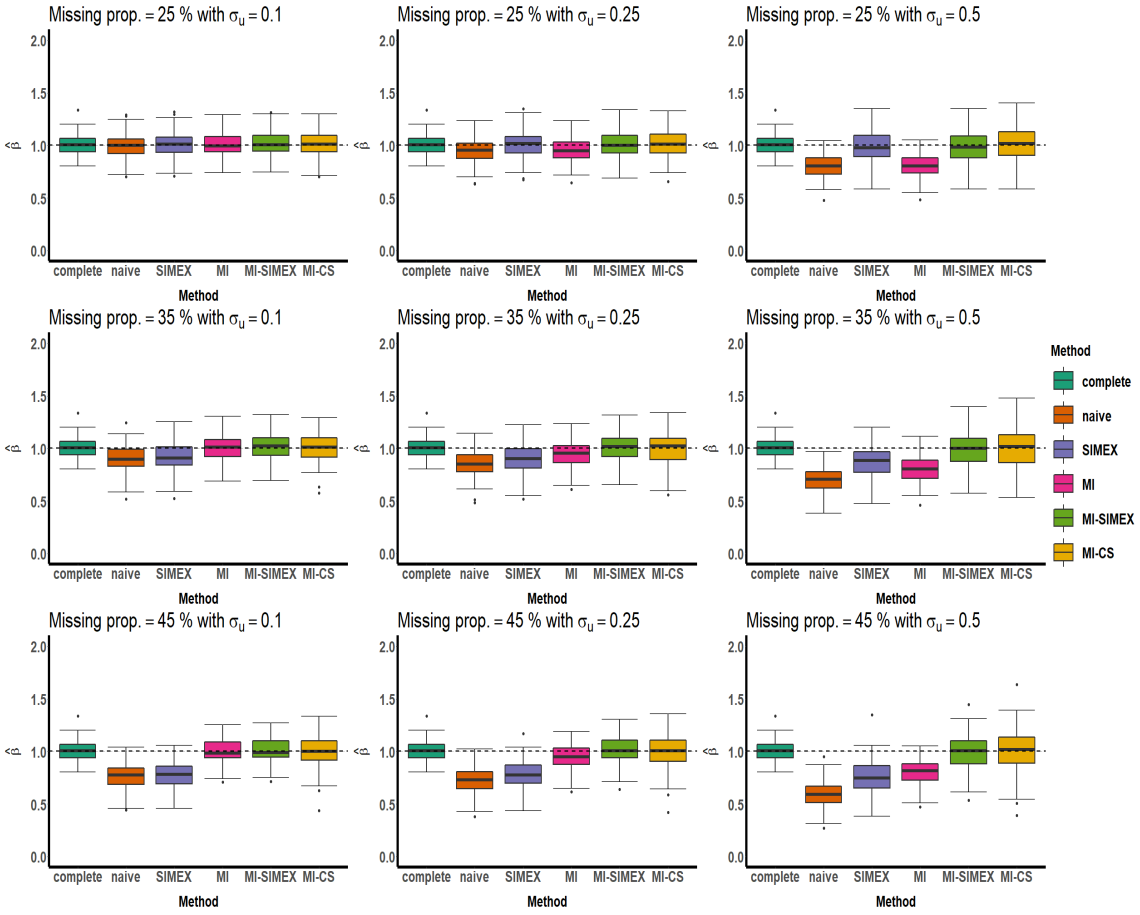


Figure 5.1. **Normal linear model: Boxplot of the slope coefficient estimates for different methods when $X \sim N(0,1)$ and $n = 100$.** Different methods are compared with the original dataset (complete, in red) based on the accuracy of their estimators as the missing proportion and the error variance increase from the top-left corner to the bottom-right. Note that the true value of the slope coefficient is 1.

MI-CS. Also, as expected, for both scenarios the bias was worst for large σ_U and higher missing proportion when using the naïve model. In summary, MI-SIMEX and MI-CS outperform the other methods in terms of the magnitude of bias for both higher values of σ_U as well as higher missing proportions where in most caese MI-SIMEX showed better performance than MI-CS, in particular, for smaller sample sizes. However, the performance of MI-CS improved as the sample size increased. Finally, the results were similar for scenarios (a) and (b).

In Figure 5.3a (and Figures A.5a–A.6a in Appendix) and Figure 5.3b (and Figures A.5b–A.6b in Appendix), we evaluated the predictive performances of the compared methods for various n under the first and the second scenario, respectively, where we plotted the MSE of the slope estimators against increasing missing proportions and increasing σ_U. As expected, for both scenarios the MSE for

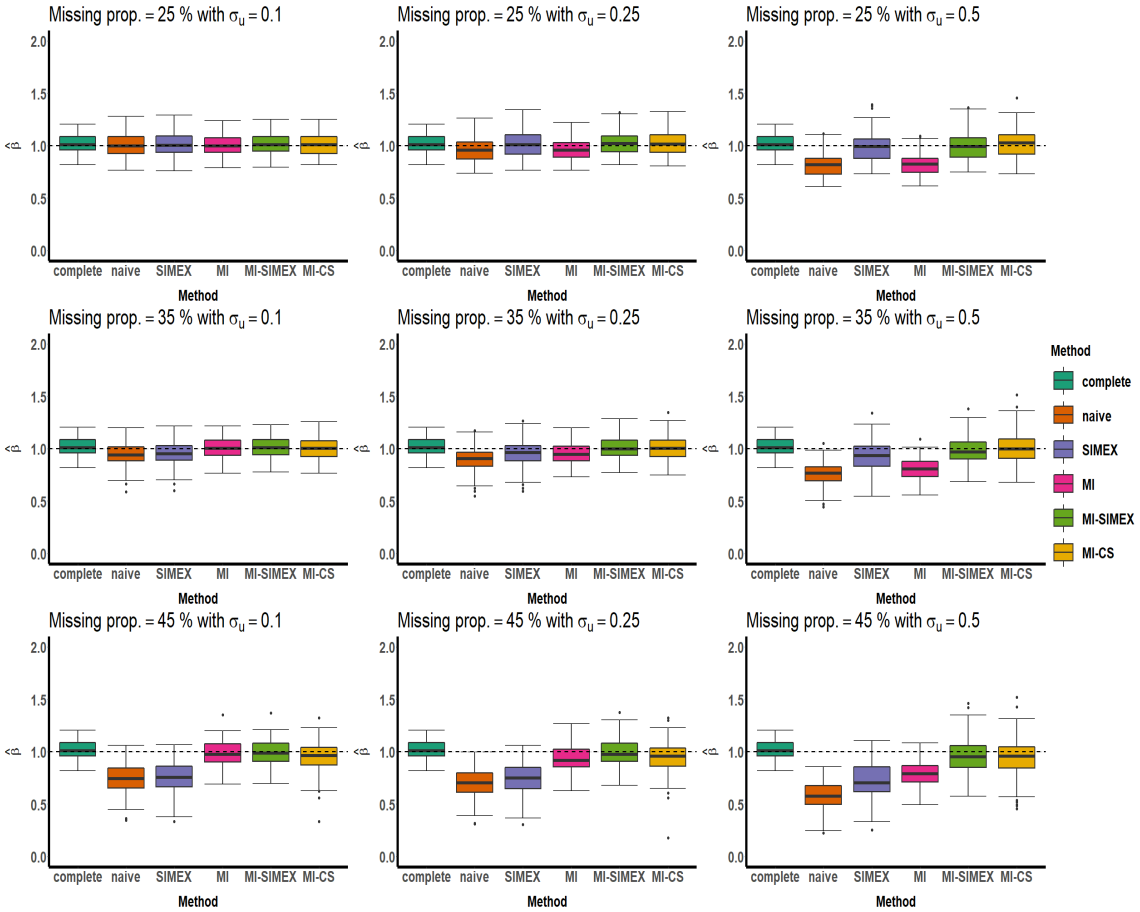


Figure 5.2. **Normal linear model: Boxplot of the slope coefficient estimates for different methods when** $X \sim U(-1, 1)$ **and** $n = 100$**.** Different methods are compared with the original dataset (complete, in mint) based on the accuracy of their estimators as the missing proportion and the error variance increase from the top-left corner to the bottom-right. Note that the true value of the slope coefficient is 1.

the naïve model had substantially increased with the missing proportion and σ_U, and was largest among the compared methods. Not surprisingly, MI reported low MSEs for small values of σ_U but started to increase with larger σ_U. MI-CS showed relatively high MSE for small sample sizes of $n = 50$, however, the predictive performance of MI-CS improved substantially as the sample size increased (*e.g.,* for $n = 100$). In summary, MI-SIMEX shows the best predictive performance among the compared methods under both scenarios and across all sample sizes.

Table 5.1. Normal linear model: Computational time (in seconds) of compared methods when $n = 100$.

complete	naive	SIMEX	MI	MI-SIMEX	MI-CS
0.002	0.001	0.326	8.200	41.084	8.263

Finally, for all methods we compared computational efficiency. Table 5.1 shows the

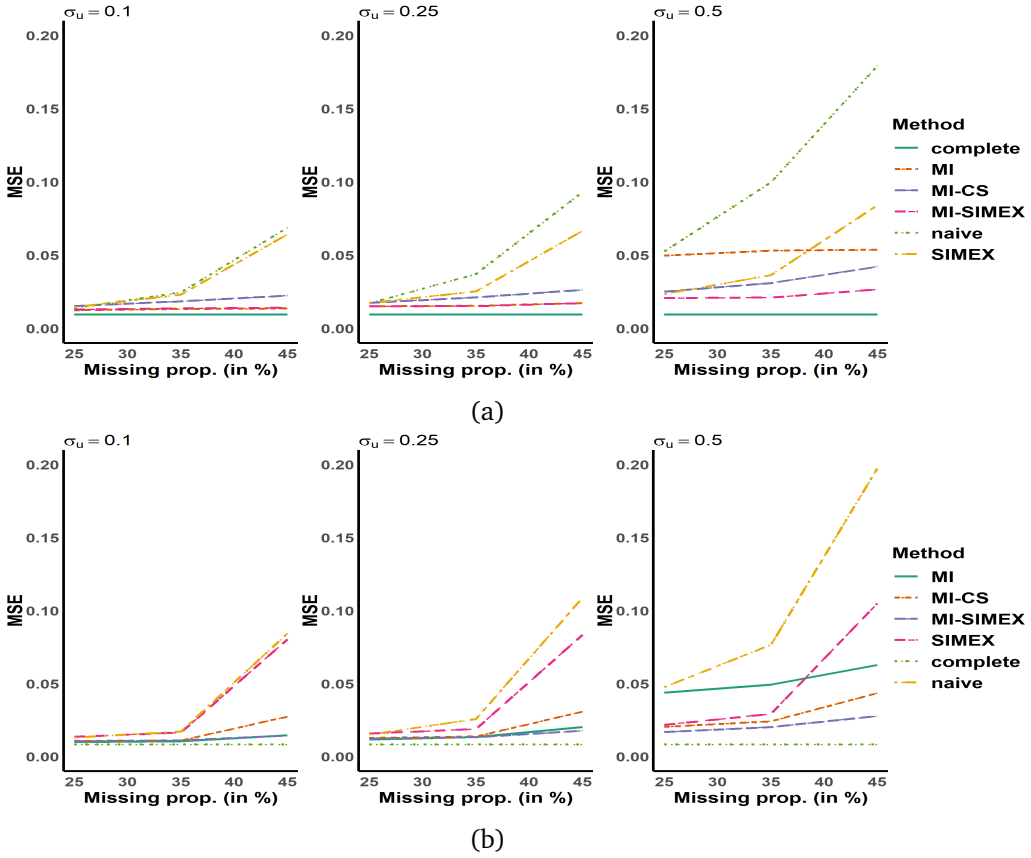


Figure 5.3. **Normal linear model: Mean squared error of the slope coefficient estimates for different methods with** $n = 100$ **when (a)** $X \sim N(0,1)$ **and (b)** $X \sim U(-1,1)$**.** Different methods are compared with the original dataset (complete, in dotted green line) for different values of error variance as the missing proportion increases.

average computational time required by each of the compared methods. Recall that the number of multiple imputations is set to $M = 100$ and the number of SIMEX resamples is set to $B = 100$ for a sample size of $n = 100$. As expected (see Section 5.2.2), MI-CS is computationally more efficient than MI-SIMEX, being about 4.97 times faster under the settings given in this simulation study. It is clear that under different parameter settings, these computational times may vary.

5.3.2 Poisson log-linear model

Next, we considered the following regression model

$$Y_i \mid X_i \overset{i.i.d.}{\sim} Poisson(\lambda_i), \quad \log \lambda_i = \beta_0 + \beta_1 X_i, \quad i = 1, \ldots, n,$$

such that the response variable is represented as counts. We used a similar simulation setting as in Section 5.3.1 for the cases where X_i was generated from (a) the

normal distribution with mean 0 and variance 1 and (b) the uniform distribution on the interval $(-1, 1)$. We only considered $n = 100$ for this simulation study. Once again we used the `simex` R-package for fitting SIMEX models but now wrote our own code for fitting multiple imputation and fitting corrected score models. The coding required to fit these models was very minimal as the corrected score for the Poisson model is provided in Nakamura (1990).

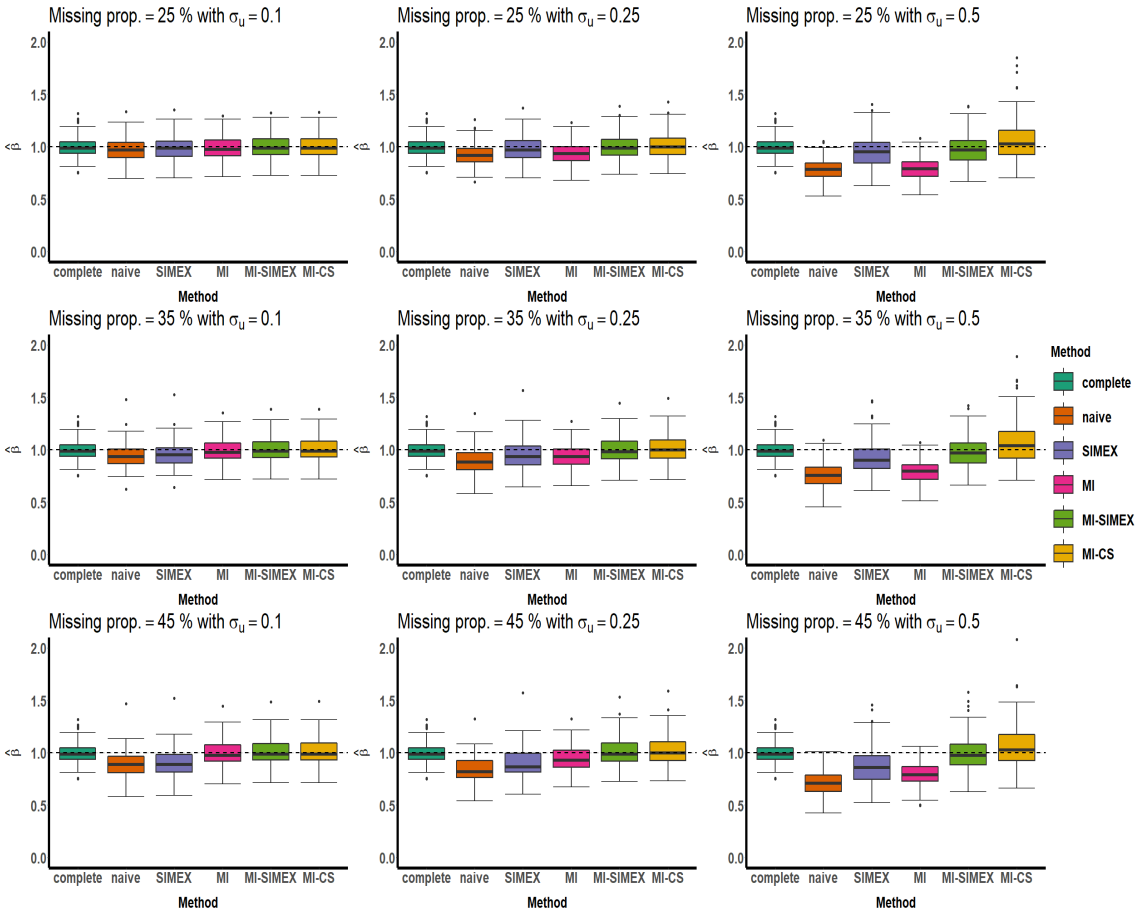


Figure 5.4. **Poisson log-linear model: Boxplot of the slope coefficient estimates for different methods when** $X \sim N(0, 1)$. Different methods are compared with the original dataset (complete, in mint) based on the accuracy of their estimators as the missing proportion and the error variance increase from the top-left corner to the bottom-right. Note that the true value of the slope coefficient is 1.

We fitted the same models as in Section 5.3.1. Figures 5.4–5.5 show parallel boxplots for the slope coefficient estimate from each fitted model for (a) and (b), respectively. Figure 5.6 shows the MSE for both (a) and (b). The results were very similar to the simulation study in Section 5.3.1 where MI-CS and MI-SIMEX outperformed the other methods in terms of the magnitude of bias and smaller MSE as the missing proportion and σ_U increased (from top-left to bottom-right).

This demonstrates that non-normal data (or at the very least, count data) can be modelled using the proposed MI-SIMEX and MI-CS, which yielded small bias compared to naïve models.

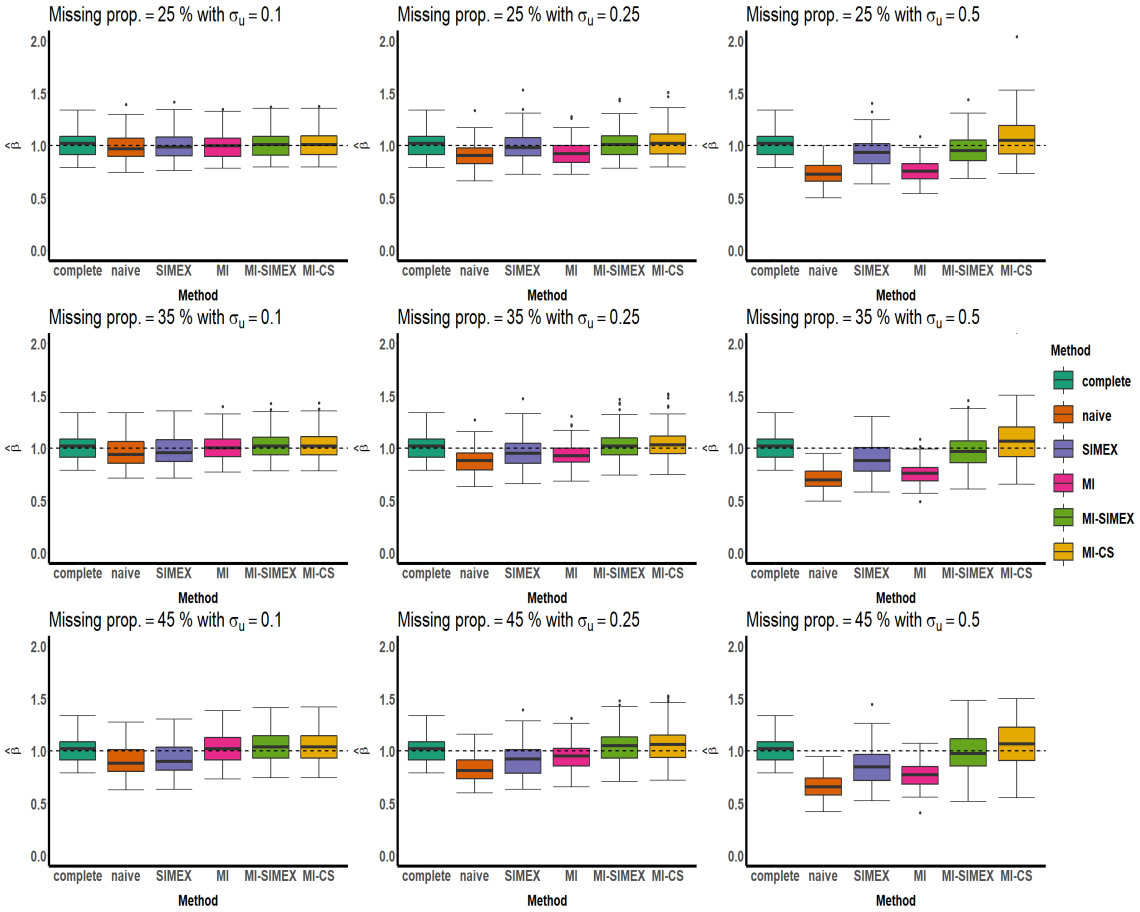


Figure 5.5. **Poisson log-linear model: Boxplot of the slope coefficient estimates for different methods when** $X \sim U(-1, 1)$. Different methods are compared with the original dataset (complete, in mint) based on the accuracy of their estimators as the missing proportion and the error variance increase from the top-left corner to the bottom-right. Note that the true value of the slope coefficient is 1.

5.4 Ozone data revisited

We fitted the same models as in our simulation studies (see Section 5.3.1) and used the same (default) setting for the `mice` and `simex` R-packages. Also, as Breiman and Friedman (1985) showed, there are no transformations required for the ozone response variable, hence we focus on the analysis on the raw data. This also makes interpretation much easier. Furthermore, we consider temperature at Sandburg as a replication of temperature measurements in order to obtain an estimate of the measurement error standard deviation at $\hat{\sigma}_U = 0.2274$. Figure 5.7a shows a

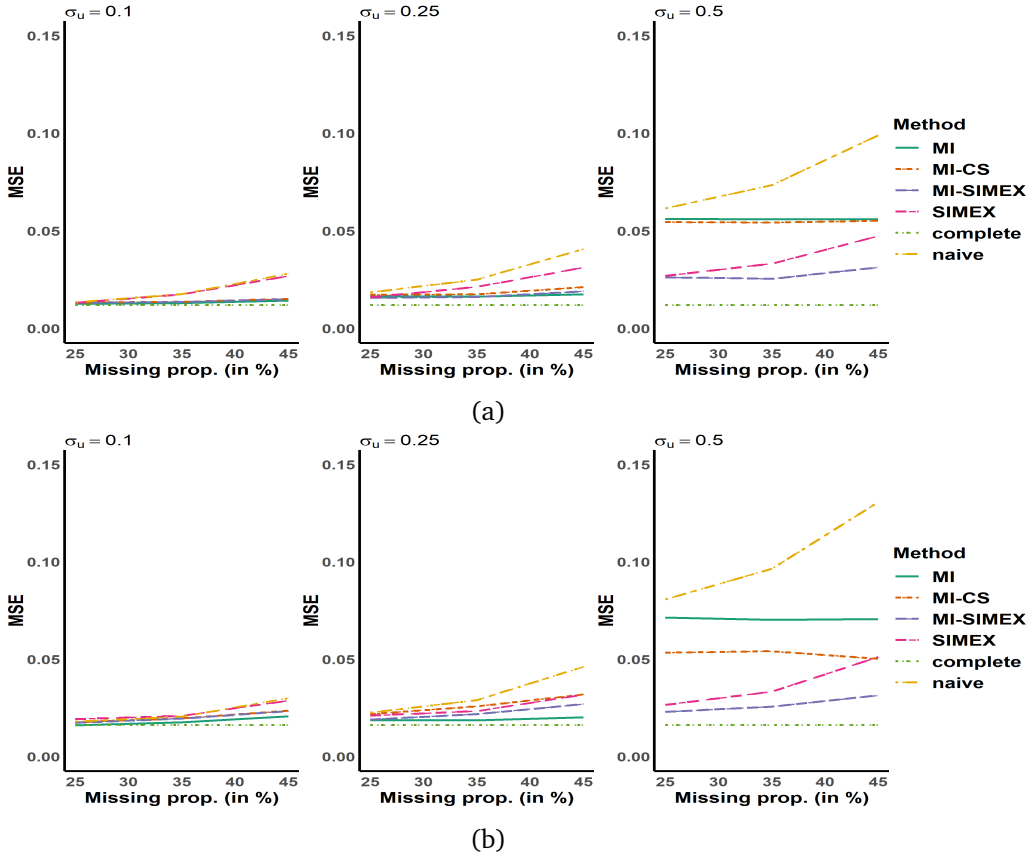
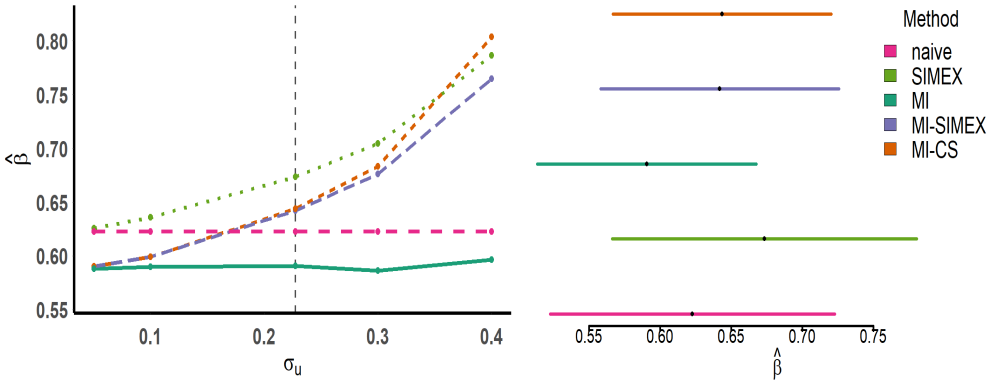


Figure 5.6. **Poisson log-linear model: Mean squared error of the slope coefficient estimates for different methods when (a)** $X \sim N(0,1)$ **and (b)** $X \sim U(-1,1)$. Different methods are compared with the original dataset (complete, in dotted green line) for different values of error variance as the missing proportion increases.

sensitivity analysis on the estimates of the regression coefficient of temperature for different values of known measurement error standard deviations (σ_U) based on various methods. First, if measurement error in the temperature covariate is ignored, it is evident that some consideration is needed to account for missing values in temperature (notice the difference in estimates between naïve and MI for $\sigma_U = 0.1$). This is likely due to the large rate of missingness. Second, as σ_U increases, the difference between measurement error model estimates and the naïve model becomes more apparent.

At $\hat{\sigma}_U = 0.2274$ (that is, the estimated measurement σ_U), the MI-SIMEX and MI-CS estimates almost coincide where they estimate this coefficient at 0.642 and 0.644 with standard errors 0.043 and 0.039, respectively. However, based on their standard errors as shown in Figure 5.7b, MI-CS seems to be slightly more efficient than MI-SIMEX for this large dataset of 366 observations. However, the spreads



(a) (b)

Figure 5.7. (a) Estimates of the slope coefficient for temperature for different methods against increasing σ_U. The grey vertical dashed line is the estimated measurement error standard deviation $\hat{\sigma}_U = 0.2274$. Notice the similarity between MI-SIMEX and MI-CS. (b) Estimates of the slope coefficient (with error bars) when $\hat{\sigma}_U = 0.2274$ for each method.

were quite similar for all models based on the given error bars.

As shown here, care needs to be taken when fitting models with climate variables as we would get different results if we do not account for both missingness and measurement error. See Foster *et al.* (2012) and Stoklosa *et al.* (2015) for a discussion on errors-in-variables for climate data modelling in the complete data case. Also, it was very easy to fit the proposed models, which with the available packages, are also computationally feasible in practice.

5.5 Discussion

Our findings suggest that combining MI with measurement error modelling methods makes it possible to deal with the combined problem of missingness and measurement error in explanatory variables in a reasonably accurate and efficient manner. In particular, MI-SIMEX showed a reasonably good performance in dealing with such combined effect on regression coefficient estimation in linear regression setting even for small sample sizes of 50 and for high missing proportions of 45%. We also investigated the case where the response variable was non-normal (count data). Again, we observed small bias and MSE compared to the naïve model fit. We note that binary data (or logistic regression models) could also be employed here using the methods proposed by Chen *et al.* (2015) and Song *et al.* (2018).

Although we primarily focused on error-contaminated explanatory variables, however, measurement error modelling methods available for error-contaminated response variables may similarly be combined with MI when missingness and measurement error both occur in the response variable. Also, this work could easily be extended to cases where one variable is subject to missingness and the other subject to measurement error by applying MI for the partially observed variable and combining it with a suitable measurement error model for the contaminated variable in a manner discussed in Section 5.2.2. Moreover, we specifically considered numerical work in this chapter and a future study could investigate the theoretical properties of the proposed methods.

Chapter 6

Discussion

6.1 Summary

Missing data problems are embedded in many research areas due to a wide variety of reasons. Perhaps, one could argue that missing data are hardly ever missed in a dataset. In Chapter 1, we introduced two datasets collected from health research studies as well as two ecological datasets where missingness appears, which we analyzed in later chapters. We also discussed the effects of missing data on statistical analysis. We found that missing data can: (1) complicate the analysis since most standard statistical analyses are designed for complete datasets, (2) cause bias in parameter estimation when missingness in not entirely due to chance, (3) cause efficiency loss due to the missing information, and (4) reduce statistical power due to the reduction in the sample size.

Two well-known missing data analysis methods are MI and MLE via EM algorithm that are often treated as distinct in the literature. For example, the difference in their performance has been the focus of several applied research studies, such as in Ho *et al.* (2001); Lynn (1982); Newman (1928); Messer and Natarajan (2008) and Lin (2010) to name but a few. However, in Chapter 2, we showed that there is a close connection between MI and MLE. In particular, there is a type of MI (improper MI) that is equivalent to a stochastic type of EM (Stochastic EM). The proper MI and the Stochastic EM algorithms proceed very similarly, the two algorithms: (0) initially, specify a model for the complete data and make a guess

on the model parameter estimates, usually based on observed data, (1) impute missing values by randomly drawing from the conditional distribution of missing data given the observed data and an assumption for missingness mechanism, (2) update estimates of unknown parameters of the specified complete data model with missing values substituted with the imputed values, (3) iterate steps 1–2 until convergence, (4) continue steps 1–2 for extra $M > 1$ times to generate multiple imputations, (5) finally, combine the results over multiple imputations in a manner that yields valid inferences about model parameters.

The key difference between a proper MI and the stochastic EM algorithm is in step 2. A proper MI algorithm treats the model parameters θ as random and estimates them in step 2 by randomly drawing from a current approximation to their posterior distribution. The underlying reason for treating θ as random is to be able to take advantage of the law of iterated variances and to sum the within and between–imputation variances in Equation (2.2) in order to reflect the uncertainty we have about the true values of the missing data in the total variance. Also, as the MI algorithm approaches its stationary distribution, the sequence of random draws for θ provides an approximation to the *mean* of its posterior. The Stochastic EM algorithm treats the model parameters θ as fixed and estimates them in step 2 by a current approximation to their MLE.

In order to reflect uncertainty we have about the true values of the missing data in the total variance, Stochastic EM takes advantage of the available methods in the ML literature, *e.g.,* the Louis' method, which are based on the missing information principle. Also, as the Stochastic EM algorithm approaches its stationary distribution, the sequence of estimates for θ provides an approximation to the MLE, which can be interpreted as the *mode* of its posterior, if θ were treated as random with flat priors.

Improper MI belongs to a class of MI that fails to be proper because it does not yield a consistent asymptotically normal estimator of θ and a weakly unbiased estimator of its asymptotic variance when based on Rubin's rule (see Section 2.2.3). This class of MI includes imputing from a model when treating θ as fixed. For example, by plugging in the MLE (Rubin, 1987, chapter 4), this leads to an asymp-

totically biased between–imputation variance and so it will incorrectly reflect the sampling variability in Rubin's rule, leading to invalid inferences. However, in the MI algorithm, if we treat θ as fixed and plug in the MLE in step 2 (improper MI) but combine the results in step 5 based on the Louis' method, this would be equivalent to the Stochastic EM, which has desirable properties and yields valid inferences. As a result, the equivalence allows us to understand MI as a stochastic EM approximation to the MLE. Therefore, it provides potential gains as it opens avenues to access likelihood-based tools and to enhance MI's performance.

In Chapter 3, by exploiting the connection between MLE and MI, we explored the application of a range of likelihood-based tools in the multiple imputation context for imputation model selection. Our findings show that we can diagnose imputation models for misspecification using standard likelihood-based information criteria such as AIC and BIC. Furthermore, we showed that BIC is consistent for selecting an imputation model given a set of competing models. We analyzed an ecological dataset as well as two health research datasets to demonstrate the method's flexibility for imputation model selection in the presence of missing data. Moreover, we demonstrated the performance of our methods on simulated data in the presence of univariate missing data as well as of multivariate missing data.

Another example demonstrating where access to likelihood-based tools improves MI's performance is in hypothesis testing. In the MI literature, hypothesis testing based on multiple imputed datasets was proposed to obtain a modified Wald test statistic (Rubin, 1987). Subsequent work by Meng and Rubin (1992) developed a pooling procedure for a likelihood ratio test with MI for nested models, using the asymptotic relationship between the Wald test and the likelihood ratio test statistics. Although this approach might work well for large samples, it can be quite cumbersome to implement in practice. In Chapter 4, using connections with maximum likelihood, we developed a StEM likelihood ratio statistic that could be directly constructed for MI. The StEM likelihood ratio statistic is easy to implement and our simulation study shows that it performs well for small samples as well as for small effect sizes. We analyzed a health research dataset to demonstrate the method's flexibility for hypothesis testing in the presence of multiple imputed data.

Another application of the equivalence between MI and StEM is where explanatory variables (or covariates) used in regression analysis are imprecisely measured in addition to containing missing values. Methods that simultaneously address both have rarely been addressed in the literature. In Chapter 5, again, by exploiting the connection between MLE and MI, we combined the likelihood-based MI with two well-known measurement error methods: SIMEX and CS. This unified approach has several appealing characteristics: the fitting procedure is easy to understand and an off-the-shelf software can be incorporated into the model fitting procedure; no calibration data or a validation subset is required; both measurement error methods are functional (*i.e.,* the distribution of the true covariates need not be specified). We demonstrated our methods on simulated data under different scenarios as well as on an ecological dataset. Our numerical work suggests that combining MI with measurement error modelling methods makes it possible to deal with the combined problem of missingness and measurement error in explanatory variables in a reasonably accurate and efficient manner.

6.2 Future work

The application of our findings may be extended to other areas where MI's performance could be improved in a likelihood based framework. One important potential extension can be in understanding how to make predictions. Currently, there is no clear guidance in the MI literature as to how to carry out prediction with multiple imputed data based on Rubin's rules except for some ad hoc suggestions (Vergouwe *et al.*, 2010; Wood *et al.*, 2015). For example, Wood *et al.* (2015) compared pooled predictions with predictions based on pooled linear predictors over multiple imputations. By distinguishing between the situations where both response and explanatory variables are partially observed and where only response variables are partially observed, they concluded that, based on their numerical work, the choice between the two approaches should be justified within the context of the prediction model: either from a second set of multiple imputations which do not include the observed response variables, or from a set of partial prediction models constructed for each potential pattern of observed ex-

planatory variables. These approaches are not straightforward and cannot be easily generalized. Thus, due to the limitations of ad hoc approaches and the lack of consensus in this area, applied researchers often feel confused how to perform prediction with MI. The equivalence bettwen MI and StEM can provide an opportunity to address this problem, since carrying out prediction in the ML literatue is a straightforward activity.

Another potential extension is in measurement error modelling when errors-in-variables can occur in any type of variable. In this thesis, we primarily focused on contaminated explanatory variables. Measurement error modelling methods available for contaminated response variables may similarly be combined with MI when missingness and measurement error both occur in the response variable. For example, in a survey analysis of population income poverty, Nicoletti *et al.* (2011) provided bounds on the poverty rate based on previous studies that deal with the combined problem of missingness and measurement errors in the response variable. Liang *et al.* (2007) considered the case where the explanatory variable is subject to measurement error but missingness occurs in the response variable in partially linear models and proposed a kernel-based imputed empirical likelihood approach to estimate regression coefficients. This work could easily be extended to the cases where one variable is subject to missingness and the other is subject to measurement error by applying MI for the partially observed variable and combining it with a suitable measurement error model for the contaminated variable in a manner discussed in Chapter 5. Other extensions include modelling overdispersion in counts (say, using the negative binomial distribution), and investigating spatial models where covariates are missing and are measured with error, see Huque *et al.* (2016). It would also be of interest to examine the theoretical properties of both SIMEX and CS methods in the missing data context.

Furthermore, we primarily focused on the StEM algorithm in this thesis. However, an alternative Monte-Carlo approximation to the EM algorithm, known as MCEM (Wei and Tanner, 1990), could similarly be used. This algorithm obtains an (approximate) MLE by averaging likelihood estimates across the multiple imputations, then maximizing. The approach can be as opposed to StEM which

maximizes on each imputation and then averages the estimates. The MCEM algorithm is more efficient than the StEM algorithm for finite sample sizes and for finite number of imputations (Nielsen, 2000). This is due to the fact that StEM loses some efficiency due to the maximize-then-average strategy. Therefore, in situations where there exists little concern for computational efficiency, it would be beneficial to apply MCEM instead of StEM without having to compromise any of the results discussed in this thesis.

Finally, for imputation model selection (see Chapter 3), we focused on using a fixed value of θ rather than a random θ. Future work could study whether similar arguments apply for random parameters. One possible adjustment to the proposed approaches when θ is random is to use Bayesian information criteria instead, such as the deviance information criterion (Spiegelhalter *et al.*, 2002). Another possible extension of Chapter 3 is to consider cases where informative missingness is assumed. Here, the probability of missingness depends on the unobserved outcome. This is a case of MNAR and is commonly approached by sensitivity analysis methods under a range of assumptions reflecting different degrees of informative missingness. An example of this is seen in longitudinal data settings with dropouts where the probability of dropout depends on the unobserved outcome. Dropouts are informative about the measured outcome. Therefore, a class of models for informative missingness can be specified as a random effects model for the primary outcome. This is combined with a model for the dropout component where the random effects are treated as covariates or with a model for time to dropout, and then, a combined likelihood is used for inference (Folmann and Wu, 1995; Bartolucci and Farcomeni, 2019). In summary, the approaches presented in Chapters 3 and 4 can be studied to investigate whether they can be used in an informative missingness setting with multiple imputed data.

Appendix A

Supplementary for Chapter 5

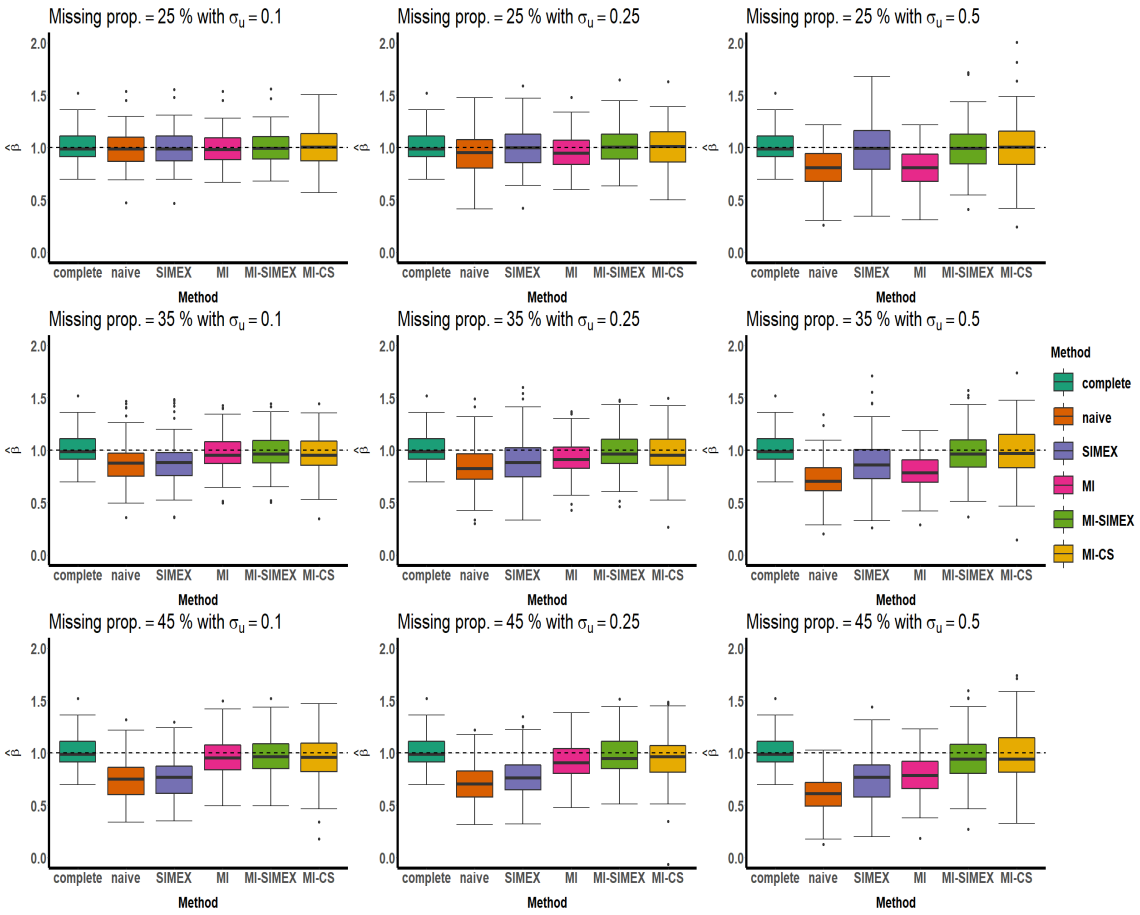


Figure A.1. **Normal linear model: Boxplot of the slope coefficient estimates for different methods when** $X \sim N(0,1)$ **and** $n = 50$**.** Different methods are compared with the original dataset (complete, in mint) based on the accuracy of their estimators as the missing proportion and the error variance increase from the top-left corner to the bottom-right. Note that the true value of the slope coefficient is 1.

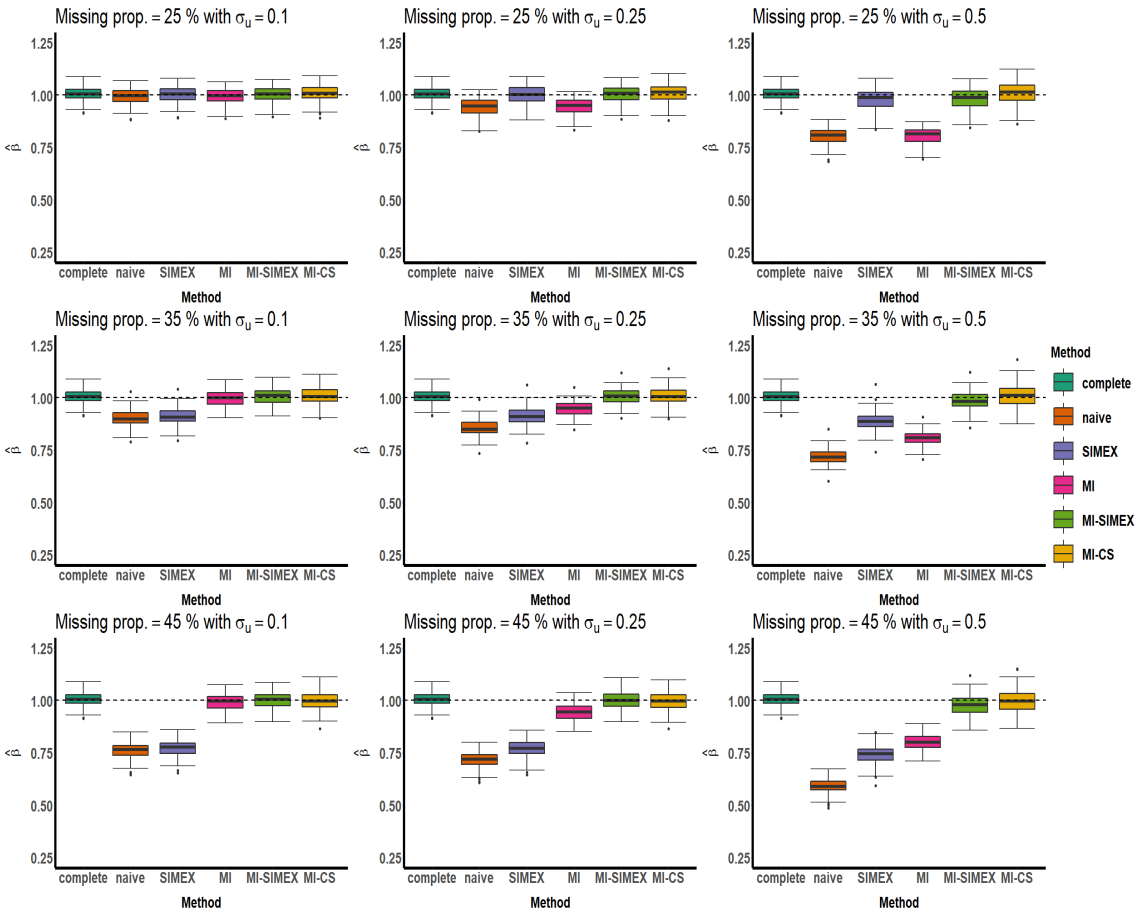


Figure A.2. **Normal linear model: Boxplot of the slope coefficient estimates for different methods when** $X \sim N(0,1)$ **and** $n = 1000$**.** Different methods are compared with the original dataset (complete, in mint) based on the accuracy of their estimators as the missing proportion and the error variance increase from the top-left corner to the bottom-right. Note that the true value of the slope coefficient is 1.

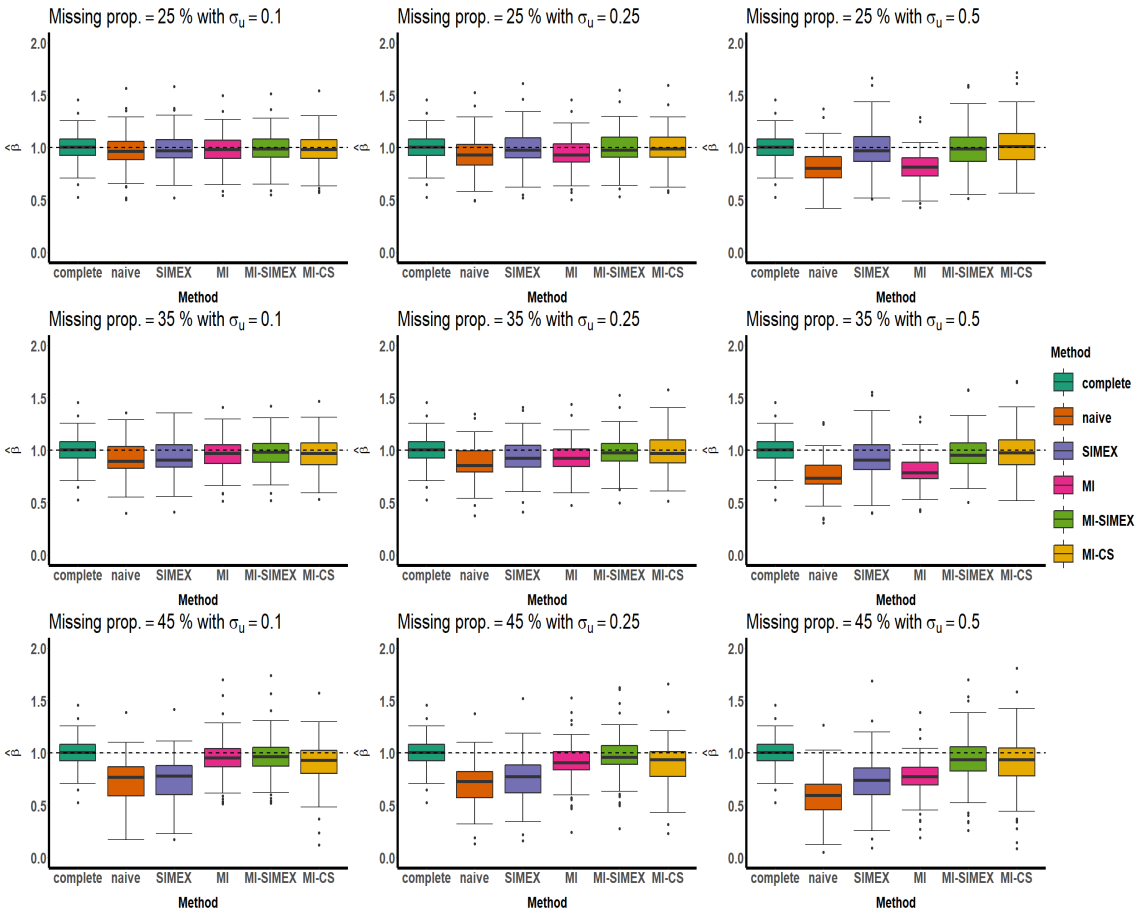


Figure A.3. **Normal linear model: Boxplot of the slope coefficient estimates for different methods when** $X \sim U(-1, 1)$ **and** $n = 50$**.** Different methods are compared with the original dataset (complete, in mint) based on the accuracy of their estimators as the missing proportion and the error variance increase from the top-left corner to the bottom-right. Note that the true value of the slope coefficient is 1.

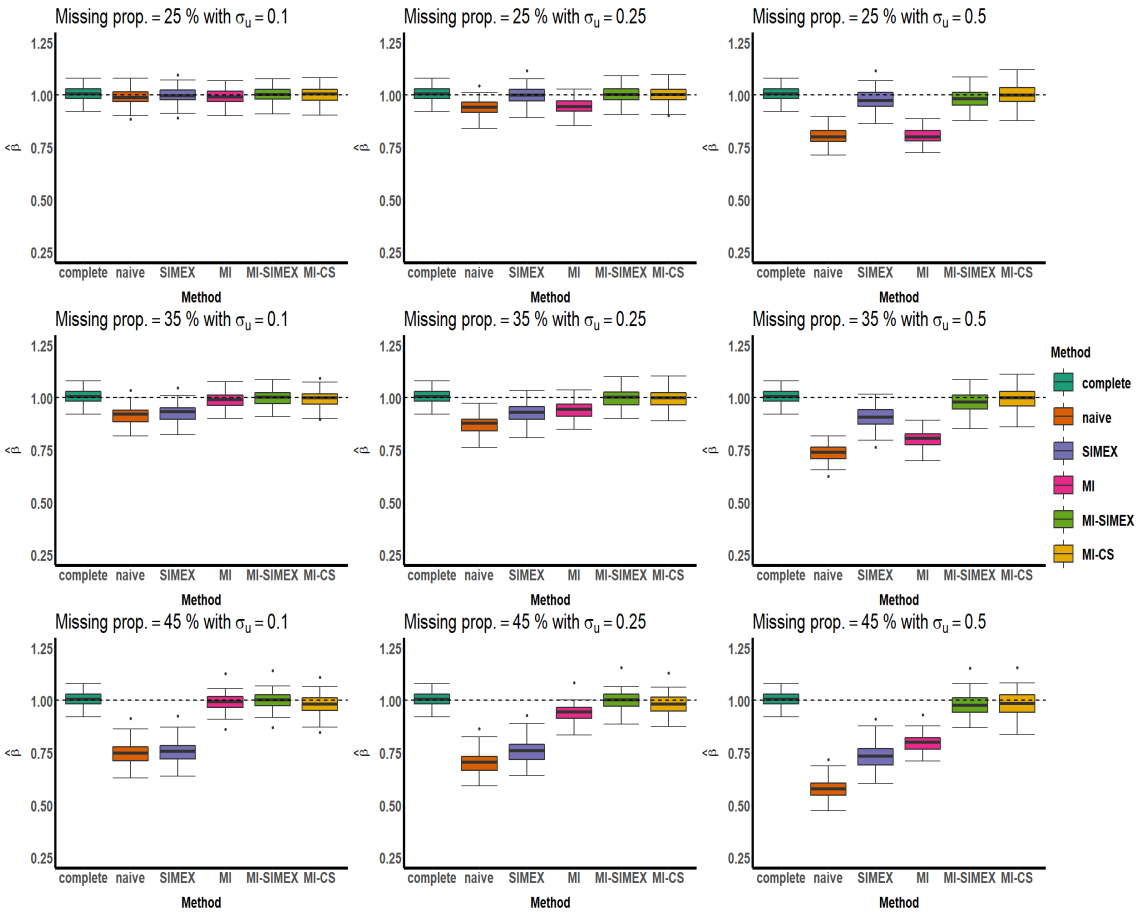
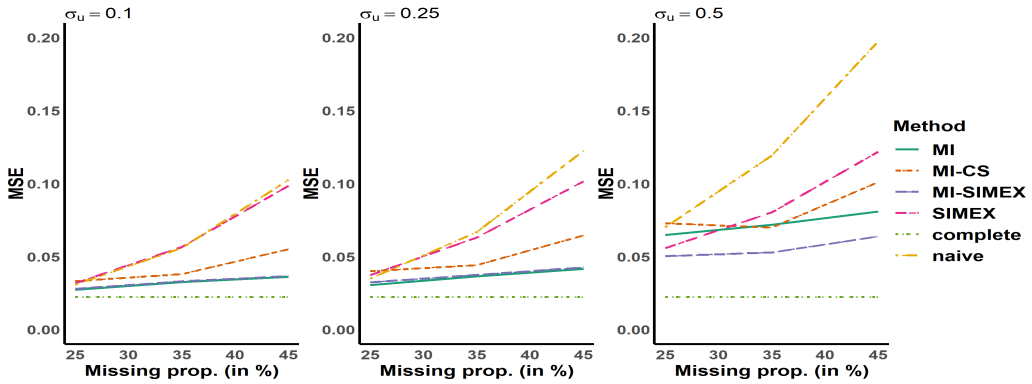
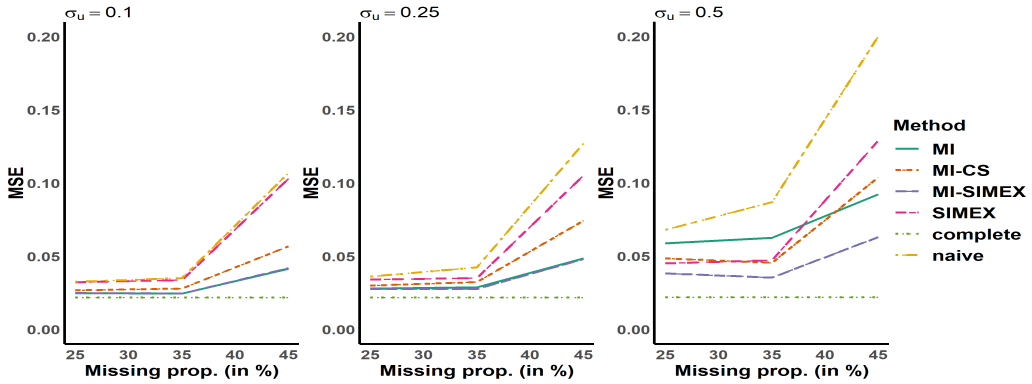


Figure A.4. **Normal linear model: Boxplot of the slope coefficient estimates for different methods when** $X \sim U(-1, 1)$ **and** $n = 1000$**.** Different methods are compared with the original dataset (complete, in mint) based on the accuracy of their estimators as the missing proportion and the error variance increase from the top-left corner to the bottom-right. Note that the true value of the slope coefficient is 1.

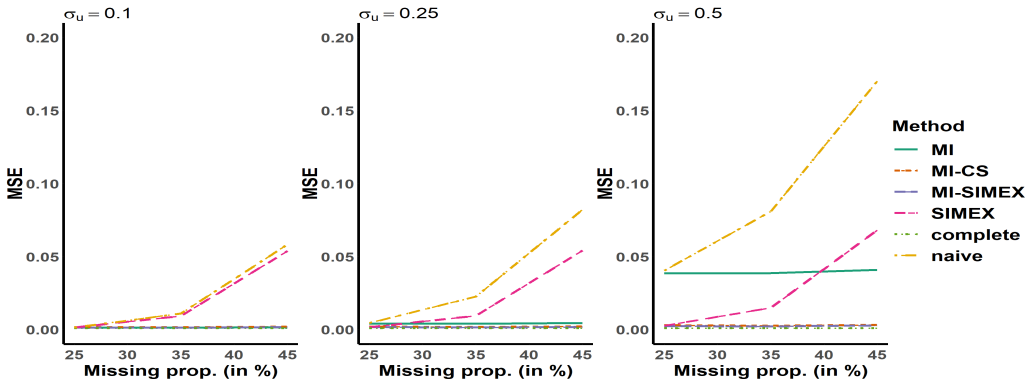


(a)

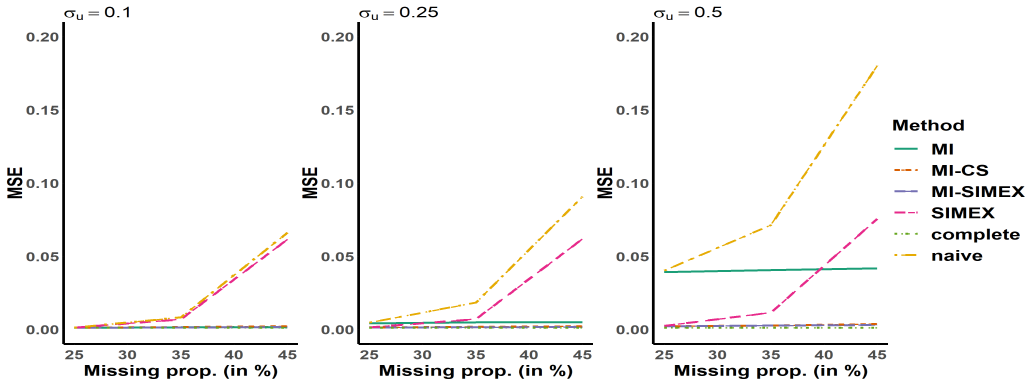


(b)

Figure A.5. **Normal linear model: Mean squared error of the slope coefficient estimates for different methods with $n = 50$ when (a) $X \sim N(0,1)$ and (b) $X \sim U(-1,1)$.** Different methods are compared with the original dataset (complete, in dotted green line) for different values of error variance as the missing proportion increases.



(a)



(b)

Figure A.6. **Normal linear model: Mean squared error of the slope coefficient estimates for different methods with $n = 1000$ when (a) $X \sim N(0,1)$ and (b) $X \sim U(-1,1)$.** Different methods are compared with the original dataset (complete, in dotted green line) for different values of error variance as the missing proportion increases.

Bibliography

Acquah H. D. G. (2010), "Comparison of Akaike information criterion (AIC) and Bayesian information criterion (BIC) in selection of an asymmetric price relationship," *Journal of Development and Agricultural Economics*, **2**(1), 001-006.

Anderson, T. W., & Goodman, L. A. (1957), "Statistical inference about Markov chains," *The Annals of Mathematical Statistics*, **28**(1), 89-110.

Andridge, R., & Thompson, K. J. (2015), "Using the fraction of missing information to identify auxiliary variables for imputation procedures via proxy pattern-mixture models," *International Statistical Review*, **83**, 472-492.

Archambeau, C., Opper, M., Shen, Y., Cornford, D., & Shawe-Taylor, J. (2007), "Variational Inference for Diffusion Processes," in *Neural Information Processing Systems*, 17-24.

Armstrong, B. (1985), "Measurement error in the generalised linear model," *Communications in Statistics – Simulation and Computation*, **14**(3), 529-544.

Arunajadai, S. G. & Rauh, V. A. (2012), "Handling covariates subject to limits of detection in regression," *Environmental and Ecological Statistics*, **19**(3), 369-391.

Barndorff-Nielsen, O. E., & Cox, D. R. (1994), *Inference and Asymptotics*, Chapman & Hall, London.

Bartolucci, F., & Farcomeni, A. (2019), "A shared-parameter continuous-time hidden Markov and survival model for longitudinal data with informative dropout," *Statistics in Medicine*, **38**(6), 1056-1073.

Blei, D. M., Kucukelbir, A., & McAuliffe, J. D. (2017), "Variational inference: A review for statisticians," *Journal of the American Statistical Association*, **112**(518), 859-877.

Biscarat, J. C., Celeux, G., & Diebolt, J. (1992), "Stochastic Versions of the EM Algorithm," Technical Report 227, Washington University, Seattle, Dept. of Statistics.

Blackwell, M., Honaker, J., & King, G. (2017), "A unified approach to measurement error and missing data: overview and applications," *Sociological Methods & Research*, **46**(3), 303-341.

Blackwell, M., Honaker, J., & King, G. (2017), "A unified approach to measurement error and missing data: Details and extensions," *Sociological Methods & Research*, **46**(3), 342-369.

Boos, D. D. & Stefanski, L. A. (2013), *Essential Statistical Inference: Theory and Methods*, Springer.

Bozdogan, H. (1987), "Model Selection and Akaike's Information Criterion (AIC): The General Theory and Its Analytical Extensions," *Psychometrika*, **52**, 345-370.

Breiman, L., & Friedman, J. H. (1985), "Estimating optimal transformations for multiple regression and correlation," *Journal of the American statistical Association*, **80**(391), 580-598.

Broniatowski, M., Celeux, G., & Diebolt, J. (1983), "Reconnaissance de Melanges de Densites par un Algorithme d'Apprentissage Probabiliste," *Data Analysis and Informatics*, **3**, 359-374.

Brooks, S., Gelman, A., Jones, G., & Meng, X. L. (Eds.) (2011), *Handbook of Markov chain Monte Carlo*, CRC press.

Buja, A., Hastie, T., & Tibshirani, R. (1989), "Linear smoothers and additive models," *The Annals of Statistics*, 453-510.

Buse, A. (1982), "The likelihood ratio, Wald, and Lagrange Multiplier tests: an expository note," *The American Statistician*, **36**, 153-157.

Buzas, J. S., & Stefanski, L. A. (1996), "A note on corrected-score estimation," *Statistics and Probability Letters*, **28**(1), 1-8.

Carpenter, J., & Kenward, M. (2012), *Multiple imputation and its application*, John Wiley & Sons.

Carpenter, J. R., Kenward, M. G., & White, I. R. (2007), "Sensitivity Analysis After Multiple Imputation Under Missing at Random: A Weighting Approach," *Statistical Methods in Medical Research*, **16**, 259-275.

Carroll, R. J., Freedman, L., & Pee, D. (1997), "Design aspects of calibration studies in nutrition, with analysis of missing data in linear measurement error models," *Biometrics*, **53**(4), 1440-1457.

Carroll, R. J., Ruppert, D., Stefanski, L. A., & Crainiceanu, C. M. (2006), *Measurement Error in Nonlinear Models: A Modern Perspective*, CRC press.

Casella, G. & Berger, R. L. (2002), *Statistical Inference, 2nd Edition.*, Pacific Grove: Duxbury.

Casella, G., & Moreno, E. (2006), "Objective Bayesian variable selection," *Journal of the American Statistical Association*, **101**(473), 157-167.

Celeux, G. & Diebolt, J. (1985), "The SEM Algorithm: A Probabilistic Teacher Algorithm Derived From the EM Algorithm for the Mixture Problem," *Computational Statistics Quarterly*, **2**, 73-82.

Celeux, G. & Diebolt, J. (1986), "Comportement Asymptotique d'un Algorithme d'apprentissage Probabiliste Pour les Melanges de Lois de Probabilite," Rapport de Recherche INRIA, 563.

Celeux, G. & Diebolt, J. (1987), "A Probabilistic Teacher Algorithm for Iterative Maximum Likelihood Estimation," in *Classification and Related Methods of Data Analysis*, ed. H. H. Bock, Amsterdam: North-Holland, 617-623.

Celeux, G., Chauveau, D., & Diebolt, J. (1996), "Stochastic versions of the EM algorithm: an experimental study in the mixture case," *Journal of Statistical Computation and Simulation*, **55**(4), 287-314.

Chen, J., Hanfelt, J. J., & Huang, Y. (2015), "A simple corrected score for logistic regression with errors-in-covariates," *Communications in Statistics-Theory and Methods*, **44**(10), 2024-2036.

Cheng, C. L., & Van Ness, J. W. (1999), *Statistical Regression with Measurement Error*, Kendall's Library of Statistics 6, Arnold, London.

Chernoff, H. (1954), "On the Distribution of the Likelihood Ratio," *The Annals of Mathematical Statistics*, **25**, 573-578.

Collins L. M., Schafer J. L., & Kam C. M. (2001), "A comparison of inclusive and restrictive strategies in modern missing data procedures," *Psychological Methods*, **6**(4), 330-351.

Cook, J. R., & Stefanski, L. A. (1994), "Simulation-extrapolation estimation in parametric measurement error models," *Journal of the American Statistical Association*, **89**(428), 1314-1328.

Dean, C. B. (1992), "Testing for overdispersion in Poisson and binomial regression models," *Journal of the American Statistical Association*, **87**(418), 451-457.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977), "Maximum Likelihood From Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society: Series B*, **39**, 1-38.

Deng, D., & Paul, S. R. (2000), "Score tests for zero inflation in generalized linear models," *The Canadian Journal of Statistics*, **28**(3), 563-570.

Diebolt, J. & Celeux, G. (1993), "Asymptotic Properties of A Stochastic EM Algorithm for Estimating Mixing Proportions," *Communications in Statistics, Series B (Stochastic Models)*, **9**, 599-613.

Diebolt, J. & Ip, E. H. S., (1996), Stochastic EM: method and application. In *Markov Chain Monte Carlo in Practice* (eds W. R. Gilks, S. Richardson and D. J. Spiegelhalter), 259-273, London: Chapman & Hall.

Dong, Y., & Peng, C. Y. J. (2013), "Principled Missing Data Methods for Researchers," *SpringerPlus*, **2**(1), 222.

Efron, B., & Tibshirani, R. (1986), "Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy," *Statistical science*, **1**(1), 54-75.

European Food Safety Authority (EFSA) (2008), "Nitrate in vegetables- Scientific Opinion on the Panel on Contaminants in the Food Chain," **689**, 1-79.

European Food Safety Authority (EFSA) (2010), "Statement on possible public health risks for infants and young children from the presence of nitrates in leafy vegetables," **8**(12), 1935.

Eugster, M. J., & Leisch, F. (2011), "Weighted and robust archetypal analysis," *Computational Statistics & Data Analysis*, **55**(3), 1215-1225.

Faes, C., Ormerod, J. T., & Wand, M. P. (2011), "Variational Bayesian inference for parametric and nonparametric regression with missing data," *Journal of the American Statistical Association*, **106**(495), 959-971.

Fay, R. E. (1991), "A Design-Based Perspective on Missing Data Variance," In *Proceedings of the 1991 Annual Research Conference*. Washington, DC: US Bureau of the Census, 429-440.

Fay, R. E. (1996), "Alternative paradigms for the analysis of imputed survey data," *Journal of the American Statistical Association*, **91**(434), 490-498.

Follmann, D., & Wu, M. (1995), "An approximate generalized linear model with random effects for informative missing data," *Biometrics*, 151-168.

Foster, S. D., Shimadzu, H., & Darnell, R. (2012), "Uncertainty in spatially predicted covariates: is it ignorable?" *Journal of the Royal Statistical Society, Series C* **61**, 637-652.

Foti, N., Xu, J., Laird, D., & Fox, E. (2014), "Stochastic Variational Inference for Hidden Markov Models," in *Neural Information Processing Systems*, 3599-3607.

Fraley, C. (1999), "On computing the largest fraction of missing information for the EM algorithm and the worst linear function for data augmentation," *Computational Statistics & Data Analysis*, **31**(1), 13-26.

Friedman, J. H., & Silverman, B. W. (1989), "Flexible parsimonious smoothing and additive modeling," *Technometrics*, **31**(1), 3-21.

Fuller, W. A. (1987), *Measurement Error Models*, Wiley Series in Probability and Mathematical Statistics, New York: Wiley.

Gamerman, D., & Lopes, H. F. (2006), *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*, Chapman and Hall/CRC.

Gelfand, A. E., & Smith, A. F. (1990), "Sampling-based approaches to calculating marginal densities," *Journal of the American Statistical Association*, **85**(410), 398-409.

Geman, S., & Geman, D. (1984), "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images," *IEEE Transactions on pattern analysis and machine intelligence*, **6**, 721-741.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson D. B., Vehtari, A., and Rubin, D. B. (2013), *Bayesian Data Analysis, 3rd Edition*. London: Chapman & Hall/CRC.

Gilks, W. R., Richardson, S., & Spiegelhalter, D. (1995), *Markov Chain Monte Carlo in Practice*, Chapman & Hall/CRC.

Graham, J. W. (2009), "Missing Data Analysis: Making it work in the real world," *Annual Review of Psychology*, **60**, 549-576.

Graham J., Olchowski A., & Gilreath T. (2007), "How many imputations are really needed? Some practical Clarifications of multiple imputation theory," *Prevention Science*, **8**(3), 206-213.

Gustafson, P. (2003), *Measurement Error and Misclassification in Statistics and Epidemiology: Impacts and Bayesian Adjustments*, Boca Raton, Florida: Chapman & Hall/CRC.

Hall, P., Penev, S., & Tran, J. (2018), "Wavelet methods for erratic regression means in the presence of measurement error," *Statistica Sinica*, **28**, 2289-2307.

Harel, O. (2007), "Inferences on missing information under multiple imputation and two-stage multiple imputation," *Statistical Methodology*, **4**(1), 75-89.

Hastie, T., Tibshirani, R., & Wainwright, M. (2015), *Statistical Learning with Sparsity: the Lasso and Generalizations*. London: Chapman & Hall/CRC.

Heitjan, D. F. & Rubin, D. B. (1991), "Ignorability and Coarse Data," *The Annals of Statistics*, **19**, 2244-2253.

Hensman, J., Fusi, N., & Lawrence, N. (2013), "Gaussian Processes for Big Data," in *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence*, Corvallis, OR: AUAI Press, 282-290.

Ho, P., Silva, M. C. M., & Hogg, T. A. (2001), "Multiple imputation and maximum likelihood principal component analysis of incomplete multivariate data from a study of the aging of port," *Chemometrics and Intelligent Laboratory Systems*, **55**(1-2), 1-11.

Huque, Md. H., Bondell, H. D., Carroll, R. J., & Ryan, L. M. (2016), "Spatial regression with covariate measurement error: A semiparametric approach." *Biometrics* **72**, 678-686.

Hurvich, C. M., & Tsai, C. L. (1989), "Regression and Time Series Model Selection in Small Samples," *Biometrika*, **76**, 297-307.

Ibrahim J.G., Lipsitz S.R. & Chen M.H. (1999), "Missing covariates in generalized linear models when the missing data mechanism is non-ignorable," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **61**(1), 173-190.

Jordan, M.I., Ghahramani, Z., Jaakkola, T.S., & Saul, L. K. (1999), "An introduction to variational methods for graphical models," *Machine learning*, **37**(2), 183-233.

Kaplan, R. (2000), *The Nothing that Is: A Natural History of Zero*, New York: Oxford University Press.

Küchenhoff, H., Mwalili, S. M., & Lesaffre, E. (2006), "A general method for dealing with misclassification in regression: the misclassification SIMEX," *Biometrics*, **62**(1), 85-96.

Kullback, S. & Leibler, R. A. (1951), "On Information and Sufficiency," *The Annals of Mathematical Statistics*, **22**, 79-86.

Lederer, W. & Kuchenhoff, H. (2006), "A short introduction to the SIMEX and MCSIMEX," *R News* **v6.4**, 26-31.

Liang, H., Wang, S., & Carroll, R. J. (2007), "Partially linear models with missing response variables and error-prone covariates," *Biometrika*, **94**(1), 185-198.

Lin, T. H. (2010), "A comparison of multiple imputation with EM algorithm and MCMC method for quality of life missing data," *Quality & Quantity*, **44**(2), 277-287.

Little, R. J. A. (1993), "Pattern-mixture Models for Multivariate Incomplete Data," *Journal of the American Statistical Association*, **88**, 125-134.

Little, R. J. A. & Rubin, D. B. (2002), *Statistical Analysis with Missing Data* (2nd ed.), New York: John Wiley and Sons.

Little, T. D., and Rhemtulla, M. (2013), "Planned missing data designs for developmental researchers," *Child Development Perspectives*, **7**(4), 199-204.

Louis, T. A. (1982), "Finding the observed information matrix when using the EM algorithm," *Journal of the Royal Statistical Society, Series B (Methodological)*, **44**(2), 226-233.

Lynn, H. S. (2001), "Maximum likelihood inference for left-censored HIV RNA data," *Statistics in Medicine*, **20**(1), 33-45.

McCullagh, P. (1998), "Maximum Likelihood Algorithms for Generalized Linear Mixed Models," *Journal of the American Statistical Association*, **92**, 162-170.

McKnight, P. E., McKnight, K. M., Sidani, S., & Figueredo, A. J. (2007), *Missing Data: A Gentle Introduction*, Guilford Press.

McLachlan, G. J., Krishnan, T., & Ng, S. K. (2004), *The EM algorithm*, Papers/Humboldt-Universität Berlin, Center for Applied Statistics and Economics (CASE).

Meng, X. L., & Rubin, D. B. (1991), "Using EM to obtain asymptotic variance-covariance matrices: The SEM algorithm," *Journal of the American Statistical Association*, **86**(416), 899-909.

Meng, X. L. & Rubin, D. B. (1992), "Performing Likelihood Ratio Tests With Multiply-imputed Data Sets," *Biometrika*, **79**, 103-111.

Messer, K., & Natarajan, L. (2008), "Maximum likelihood, multiple imputation and regression calibration for measurement error adjustment," *Statistics in Medicine*, **27**(30), 6332-6350.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953), "Equation of state calculations by fast computing machines," *The Journal of Chemical Physics*, **21**(6), 1087-1092.

Metropolis, N., & Ulam, S. (1949), "The Monte Carlo method," *Journal of the American Statistical Association*, **44**(247), 335-341.

Miller, A. (2002), *Subset Selection in Regression*, CRC Press.

Miller M., Burch T. A., Bennett P. H. & Steinber A. G. (1965), "Prevalence of diabetes mellitus in American Indians – results of glucose tolerance tests in Pima Indians of Arizona," *Diabetes*. Alexandria: Amer Diabetes Assoc, **14**(7), 439-440.

Molenberghs, G., & Kenward, M. (2007), *Missing Data in Clinical Studies* (Vol. 61), John Wiley & Sons.

Muggeo, V. M., & Lovison, G. (2014), "The 'three plus one' likelihood-based test statistics: unified geometrical and graphical interpretations," *The American Statistician*, **68**(4), 302-306.

Nakamura, T. (1990), "Corrected score function for errors-in-variables models: Methodology and application to generalized linear models," *Biometrika*, **77**(1), 127-137.

Ng, S. K., Krishnan, T., & McLachlan, G. J. (2012), "The EM algorithm," In Handbook of *computational statistics*, Springer, Berlin, Heidelberg, 139-172.

Newman, D. A. (2003), "Longitudinal modeling with randomly and systematically missing data: A simulation of ad hoc, maximum likelihood, and multiple imputation techniques," *Organizational Research Methods*, **6**(3), 328-362.

Neyman, J., & Pearson, E. S. (1928), "On the use and interpretation of certain test criteria for purposes of statistical inference," *Biometrika*, **20**, 175-240.

Ng, S. K. & McLachlan, G. J. (2003), "On the choice of the number of blocks with the incremental EM algorithm for the fitting of normal mixtures," *Statistics and Computing*, **13**, 45-55.

Nicoletti, C., Peracchi, F., & Foliano, F. (2011), "Estimating income poverty in the presence of missing data and measurement error," *Journal of Business and Economic Statistics*, **29**(1), 61-72.

Nielsen S. F. (2000), "The Stochastic EM Algorithm: Estimation and Asymptotic Results," *Bernoulli*, **6**, 457-489.

Nishii, R. (1984), "Asymptotic Properties of Criteria for Selection of Variables in Multiple Regression," *The Annals of Statistics*, **12**, 758-765.

Novick, S. J., & Stefanski, L. A. (2002), "Corrected score estimation via complex variable simulation extrapolation," *Journal of the American Statistical Association*, **97**(458), 472-481.

Ormerod, J. T., & Wand, M. P. (2010), "Explaining Variational Approximations," *The American Statistician*, **64**, 140-153.

Penev, S., & Raykov, T. (1993), "On choosing estimators' in a simple linear errors-in-variables model," *Communications in Statistics–Theory and Methods*, **22**(9), 101-115.

Pregibon, D. (1980), "Goodness of link tests for generalized linear models," *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, **14**, 15-23.

Peng, C. Y. J., Harwell, M., Liou, S. M., & Ehman, L. H. (2006), "Advances in missing data methods and implications for educational research," *Real Data Analysis*, 31-78.

Quijano, L., Yusà, V., Font, G., McAllister, C., Torres, C., & Pardo, O. (2017), "Risk assessment and monitoring programme of nitrates through vegetables in the Region of Valencia (Spain)," *Food and Chemical Toxicology*, **100**, 42-49.

R Development Core Team. (2019). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3–900051–07–0. Available at http://www.r-project.org.

Raghunathan, T. E. (2004), "What do we do with missing data? Some options for analysis of incomplete data," *Annual Review of Public Health*, **25**, 99-117.

Rao, C. R. (1948), "Large sample tests of statistical hypotheses concerning several parameters with application to problems of estimation," *Proceedings of the Cambridge Philosophical Society*, **44**, 50-57.

Rao, C. R. (1973), *Linear Statistical Inference and its Applications, 2nd Edition*, New York: John Wiley & Sons.

Rao, C. R. (2005), "Score test: historical review and recent developments," In *Advances in Ranking and Selection, Multiple Comparisons, and Reliability, (Eds., N. Balakrishnan, N. Kannan, H.N. Nagaraja)*, Chapter 1, p. 3-20, Birkhhaeuser, Boston.

Rayner, J. C. W. (1997), "The asymptotically optimal tests," *Statistician*, **46**, 337-346.

Ripley B. D. (1996), *Pattern Recognition and Neural Networks*. Cambridge: Cambridge University Press.

Rubin, D. B. (1976), "Inference and Missing Data," *Biometrika*, **63**, 581-592.

Rubin, D. B. (1987), *Multiple Imputation for Nonresponse in Surveys*, New York: Wiley.

Rubin, D. B. (1996), "Multiple Imputation After 18+ Years," *Journal of the American Statistical Association*, **91**, 473-489.

Schafer, J. L. (1997), *Analysis of Incomplete Multivariate Data*, London: Chapman & Hall.

Schafer, J. L. (1999), "Multiple imputation: a primer," *Statistical Methods in Medical Research*, **8**(1), 3-15.

Schafer, J. L., & Graham, J. W. (2002), "Missing data: our view of the state of the art," *Psychological Methods*, **7**(2), 147.

Schomaker, M., & Heumann, C. (2014), "Model selection and model averaging after multiple imputation," *Computational Statistics & Data Analysis*, **71**, 758-70.

Schwarz, G. (1978), "Estimating the Dimension of A Model," *Annals of Statistics*, **6**, 461-464.

Seaman, S. R. & White, I. R. (2011), "Review of inverse probability weighting for dealing with missing data," *Statistical Methods in Medical Research*, **22**(3), 278-295.

Sen, P. K., & Singer, J. M. (1994), *Large Sample Methods in Statistics: An Introduction With Applications* (Vol. 25), CRC Press.

Shah, R. D., & Bühlmann, P. (2018), "Goodness-of-fit tests for high dimensional linear models," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **80**(1), 113-135.

Shen, C-W., & Chen, Y-H. (2012), "Model selection for generalized estimating equations accommodating dropout missingness," *Biometrics*, **68**, 1046-1054.

Shen, C-W. & Chen, Y-H. (2016), "Model selection for marginal regression analysis of longitudinal data with missing observations and covariate measurement error," *Biostatistics*, **16**(4), 740-753.

Shibata, R. (1976), "Selection of the Order of An Autoregressive Model by Akaike's Information Criterion," *Biometrika*, 63, 117-126.

Smith J. W., Everhart J. E., Dickson W. C., Knowler W. C. & Johannes R. S. (1988), "Using the ADAP learning algorithm to forecast the onset of diabetes mellitus," *Proc Annu Symp Comput Appl Med Care*, IEEE Computer Society Press, 261-265.

Song, X., & Wang, C. Y. (2018), "GMM nonparametric correction methods for logistic regression with error-contaminated covariates and partially observed instrumental variables," *Scandinavian Journal of Statistics*, in press.

Spiegelhalter D. J., Best N. G., Carlin B. P., & Van Der Linde A. (2002), "Bayesian measures of model complexity and fit," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **64**(4), 583-639.

Stefanski, L. A. (1989), "Unbiased estimation of a nonlinear function a normal mean with application to measurement error for models," *Communications in Statistics-Theory and Methods*, **18**(12), 4335-4358.

Stefanski, L. A. (2000), "Measurement error models," *Journal of the American Statistical Association*, **95**(452), 1353-1358.

Stefanski, L. A., & Carroll, R. J. (1985), "Covariate measurement error in logistic regression," *The Annals of Statistics*, **13**(4), 1335-1351.

Stefanski, L. A., & Cook, J. R. (1995), "Simulation-extrapolation: the measurement error jackknife," *Journal of the American Statistical Association*, **90**(432), 1247-1256.

Sterne, J. A., White, I. R., Carlin, J. B., Spratt, M., Royston, P., Kenward, M. G., Wood A. M., Carpenter, J. R. (2009), "Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls," *BMJ*, **338**, b2393.

Stoklosa, J., Gibb, H., & Warton, D. I. (2014), "Fast forward selection for generalized estimating equations with a large number of predictor variables," *Biometrics*, **70**, 110-120.

Stoklosa, J., Daly, C., Foster, S. D., Ashcroft, M. B., & Warton, D. I. (2015), "A climate of uncertainty: accounting for error in climate variables for species distribution models," *Methods in Ecology and Evolution*, **6**(4), 412-423.

Stoklosa, J., Lee, S-H., & Hwang, W. H. (2019), "Closed-population Capture-recapture Models With Measurement Error and Missing Observations in Covariates," *Statistica Sinica***29**, 589-610.

Tanner, M. A and Wong, W. H. (1987), "The Calculation of Posterior Distributions by Data Augmentation," *Journal of the American Statistical Association*, **82**, 528-540.

Tierney, L. (1994), "Markov chains for exploring posterior distributions," *The Annals of Statistics*, **22**(4), 1701-1728.

Tierney L. (1999), "Exploring posterior distributions using Markov chains," In *Computer Science and Statistics: Proc. 23rd Symp. Interface (ed. E. Keramidas)*. Fairfax Station: Interface Foundation, 1563-1570.

Van Buuren, S. (2012), *Flexible Imputation of Missing Data*. Boca Raton: Chapman & Hall CRC press.

Buuren, S. van, & Groothuis-Oudshoorn, K. (2011), "MICE: Multivariate imputation by chained equations," in *R. Journal of Statistical Software*, **45**(3), 1-67.

Vaeth, M. (1985), "On the use of Wald's test in exponential families," *International Statistical Review*, 199-214.

Vergouwe, Y., Royston, P., Moons, K. G., & Altman, D. G. (2010), "Development and validation of a prediction model with missing predictor data: a practical approach," *Journal of Clinical Epidemiology*, **63**(2), 205-214.

von Hippel, P. T. (2013), "The bias and efficiency of incomplete-data estimators in small univariate normal samples," *Sociological Methods and Research*, **42**, 531-558.

Wald, A. (1943), "Tests of statistical hypotheses concerning several parameters when the number of observations is large," *Transactions of the American Mathematical Society*, **54**, 426-482.

Wang, C. Y., Huang, Y., Chao, E. C., & Jeffcoat, M. K. (2008), "Expected estimating equations for missing data, measurement error, and misclassification, with application to longitudinal nonignorable missing data," *Biometrics*, **64**(1), 85-95.

Wang, N., & Robins, J. M. (1998), "Large-sample Theory for Parametric Multiple Imputation Procedures," *Biometrika*, **85**, 935-948.

Wang, M., Sun, X., & Lu, T. (2015), "Bayesian structured variable selection in linear regression models," *Computational Statistics*, **30**(1), 205-229.

Warton, D. I. (2008), "Which Wald statistic? Choosing a parameterisation of the Wald statistic to maximise power in k-sample generalised estimating equations," *Journal of Statistical Planning and Inference*, **138**, 3269-3282.

Wei, G. C. G., & Tanner, M. A. (1990), "A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms," *Journal of the American Statistical Association*, **85**(411), 699-704.

White, I. R., Royston, P., & Wood, A. M. (2011), "Multiple imputation using chained equations: issues and guidance for practice," *Statistics in Medicine*, **30**(4), 377-399.

Wood A. M., Royston P., & White I. R. (2015), "The estimation and use of predictions for the assessment of model performance using large samples with multiply imputed data," *Biometrical Journal*, **57**(4), 614-632.

Orchard, T., Woodbury, M. A. (1972), "A missing information principle: theory and applications," in *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Vol. 1: Theory of Statistics*, p. 697-715, University of California Press, Berkeley, California. Available at https://projecteuclid.org/euclid.bsmsp/1200514117.

Woolrich, M., Behrens, T., Beckmann, C., Jenkinson, M., & Smith, S. (2004), "Multilevel Linear Modeling for fMRI Group Analysis using Bayesian Inference," *Neuroimage*, **21**, 1732-1747.

Wilks, S. S. (1938), "The Large-sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses," *The Annals of Mathematical Statistics*, **9**, 60-62.

Yi, G. Y., Ma, Y., & Carroll, R. J. (2012), "A functional generalized method of moments approach for longitudinal studies with missing responses and covariate measurement error," *Biometrika*, **99**(1), 151-165.